Nizomjon Jumaniyazov Baxtiyorovich

DOCTORAL DISSERTATION

NUMERICAL RESOLUTION OF FOKKER-PLANCK TYPE KINETIC EQUATIONS

Department of Applied Mathematics
Faculty of Mathematics

Santiago de Compostela

2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Doctoral Dissertation

Numerical Resolution of Fokker-Planck Type Kinetic Equations

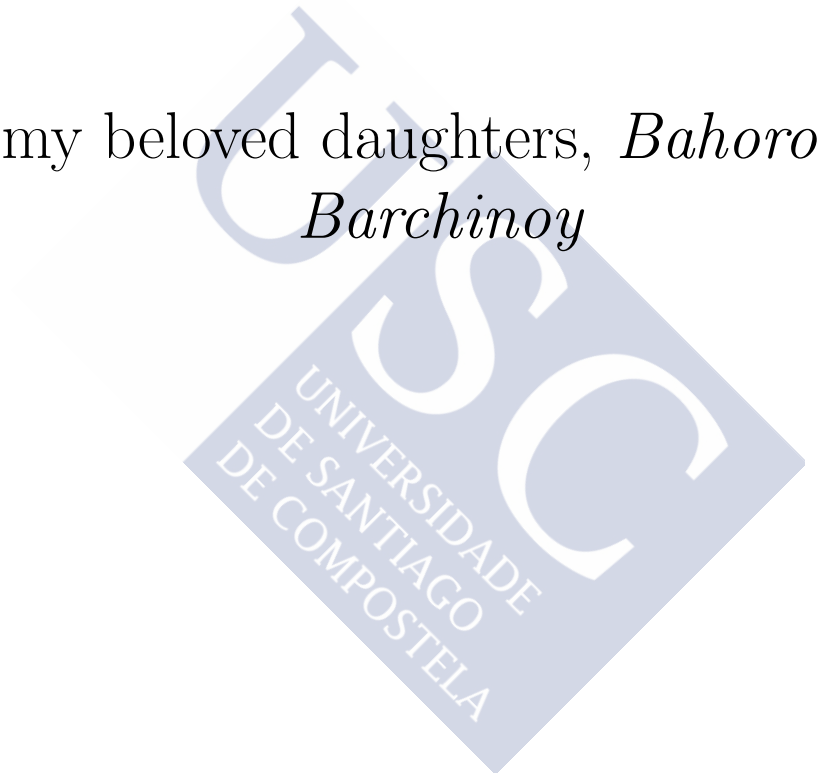
Author:
Nizomjon Jumaniyazov

Supervisor:
Óscar López Pouso

September 19, 2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

To my beloved daughters, *Bahoroy* and *Barchinoy*



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Don Óscar López Pouso, profesor titular de universidad en el Departamento de Matemática Aplicada de la Universidad de Santiago de Compostela, informa que la memoria titulada

NUMERICAL RESOLUTION OF FOKKER-PLANCK TYPE KINETIC EQUATIONS

fue realizada bajo su dirección por Don Nizomjon Jumaniyazov Baxtiyorovich, estimando que el interesado se encuentra en condiciones de optar al grado de Doctor, por lo que solicita que sea admitida a trámite para su lectura y defensa pública.

Santiago de Compostela, a 19 de septiembre de 2017.

Fdo.: Nizomjon Jumaniyazov

Fdo.: Óscar López Pouso

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Acknowledgements

First of all, I would like to say my tremendous thanks and gratitude to my best lifelong friend and colleague, *Umid Karimov*, who highly supported me coming here, to the University of Santiago de Compostela for my doctoral research. For the last 15 years, he has always been one of the best people around me, encouraging to go ahead and finish the research.

Especially, I would like to express my special and deep appreciation and thankfulness to my supervisor Professor *Óscar López Pouso*, who has been endorsing me within the research process for almost the last 10 years. It is very kind of him to encourage me to do research under his supervision and to grow as a research scientist. His support on both research and on my career has been invaluable. He always greatly enlightens me to complete my research. All along of this, I cannot thank him enough. I am forever grateful.

A special thanks to *my family*. Words cannot express how grateful I am to *my mother*, and *father* for all of the giving ups that you have made on my behalf. Your prayer for me was what sustained me thus far. I thank *my father*, for letting me through all the difficulties. I have experienced your guidance day by day. You are the one who let me finish my degree.

I would also like to give a heartfelt, special thanks to Professor *Aknazar Khasanov*, who is the founder of the mathematical school devoted to studying spectral theory of differential equations, my first research director at Urgench State University (UrSU). During my doctoral research at UrSU, he encouraged and led me to develop personally and professionally. He greatly inspired to resume my doctoral research abroad. His academic support, input and personal cheering are greatly appreciated. Thank you.

I do want to utter my thanks to my best friend from my school time *Rashid Adinayev* for being always ready to support me in any kind of situtation. My next acknowledgements go to one of my best Spanish friends, *Jorge Albella*, who has always been in a friendly relation and given a tremendous helping hand no matter the task or circumstance since the first day of my mobility program. Thank you, buddies, for everything you have done for me.

Finally, I must acknowledge with tremendous and deep thanks my teachers of Mathematics and Physics, *Otabek Egamberdiyev* and *Maqsud Otajonov*, who laid down early foundations for this thesis by highly motivating and teaching me to become a university student.

Santiago de Compostela, September 19, 2017
Nizomjon Jumaniyazov Baxtiyorovich

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Preface

The Fokker-Planck equation (FPE) is a partial differential equation that describes the time evolution of the probability density function of the velocity of a particle under the influence of drag forces and random forces, as in Brownian motion. The equation can be generalized to other observables as well. It is named after Adriaan Fokker and Max Planck and is also known as the Kolmogorov forward equation, after Andrey Kolmogorov, who independently discovered the concept in 1931.

The Fokker-Planck equation is often used to approximate the description of particle transport processes with highly forward-peaked scattering. Pomraning has shown that if the physical scattering kernel is sufficiently dominated by small-angle scattering, then the Fokker-Planck equation is an asymptotic approximation to the linear Boltzmann equation.

The main purpose of this research work is to consider a new finite difference method and an iterative method to solve the Fokker-Planck equation when the angular flux depends, in general, on three variables; spatial, polar and azimuthal variables. The work is organized as follows:

In Introduction, a brief information on the problem to be considered and the main data related to the problem are presented. Morever, similar problems studied by different authors are cited to mention differences and similarites. Furthermore, the conditions posed for the problem to be well-defined are introduced.

Chapter 1: Relationship with the general 3D FPE. Physical interpretation. Along this Chapter, relationship between general 3D and 1D Fokker-Planck equations is learned. Under some assumptions general 3D Fokker-Planck equation is reduced to 1D equation. In this sense, as a first assumption, the problem is supposed to be steady and later Fourier techniques and energy discretizations are exploited which lead the problem to 1D problem.

Chapter 2: **Analysis of particular case**. Throughout this Chapter, a special example, the case without diffusion is considered to analyse regularity and continuity of function ψ which is the solution to the problem. Continuity analyse contains up to the class k.

Chapter 3: Direct method. As the work is targeted on solving the Fokker-Planck equation with different methods, we mainly consider in this Chapter direct method which is later seen in Chapter 6 more advantageous than iterative one. The direct method is essentially based on Crank-Nicolson method which contains both implicit and explicit schemes.

We derive a numerical scheme, which includes odd and even schemes meaning when the number of μ-nodes is odd and even respectively. When the number of μ-nodes is odd, the number of equations does not coincide with that of unknowns, and a correction has to be done. Following the numerical method, which preserves the order of 2 in both μ and z variables, we obtain a system of equations. In order to check if the method works, we consider two problems in [21], which leads us to conclude that our method works. As to the order, we cannot say anything as nothing is said about the order in abovementioned paper. Morever, we focus on the different problems with exact solutions which might yield to say about the order. Appropriate tables are presented to show details obtained.

Chapter 4: MATLAB® implementation. This Chapter deals mainly with coding the numerical schemes derived in Section 3.1 in MATLAB®. From a programming point of view, a way for converting a matrix into a vector, which is the solution of the linear system to be composed by the equations in Section 3.1, plays crucial role in the code. This is done with the help of a bijective mapping, called *pointer*. Making use of the pointer, we assign the equations in the system which is finally solved easily using the MATLAB® command "\". Moreover, some examples on how to use the MATLAB® command `sparse` are presented. As the solution of the linear system posseses a lot of zeros, this command makes the code work faster. MATLAB® commands `kron` and `repmat` are very handful in solving the system as well. Using commands mentioned above, we present codes for each set of equations *(4.1)–(4.6)*.

Chapter 5: Iterative method. This Chapter is devoted to solving the same problem we consider in this work with an iterative method. For solving the problem with the iterative method, the odd scheme, which is more profitable than the even scheme is used. From a computational time viewpoint, the iterative method could be better, but according to some numerical experiments we have carried out, the direct approach turns out to be better.

Chapter 6: Direct method vs iterative method. In this Chapter, some examples are considered with both direct and iterative methods to compare their usability. Proper tables are presented to see the differences of the results obtained by means of both methods.

Chapter 7: Azimuthal angle dependend problem. This Chapter is devoted to the case the dependence on azimuthal angle θ is not neglected. As in Section 1.2, Fourier technique is employed to split the problem into a collection of θ-independent problems. In this case the absortion coefficient becomes singular, which requires to modify the odd scheme described in Section 4.1. Some numerical experiments are carried out using modified odd scheme and they show once more second order of convergence.

Contents

Preface IX

Contents XII

Introduction 1

1. Derivation of 1D FPE from the general 3D FPE. 5
1.1. Introduction . . . 6
1.2. Physical domain . . . 9
1.3. Fourier techniques . . . 9
1.4. Simplifications . . . 12
1.4.1. First and second simplifications . . . 12
1.4.2. Third and fourth simplifications . . . 13

2. Analysis of a particular case 15
2.1. Trivial problem: case without diffusion . . . 16
2.2. A particular example . . . 16
2.2.1. Regularity analysis of function ψ . . . 17
2.2.2. Continuity analysis . . . 18

3. Direct method 23
3.1. The numerical scheme . . . 24
3.1.1. Derivation of the numerical scheme . . . 26
3.1.2. Description of the numerical scheme . . . 27
3.2. Numerical results . . . 29
3.2.1. Problems with known regular solutions . . . 29
3.2.2. Examples from Kim and Tranquilli [21]. . . . 32
3.3. Order and order* . . . 38

4. MATLAB® implementation 43
4.1. Description of the odd scheme . . . 44
4.2. The MATLAB® command `sparse` . . . 52
4.3. Defining the matrix in the code . . . 56
4.3.1. Code for Equations *(4.1)* . . . 58
4.3.2. Code for Equations *(4.4)* . . . 59
4.3.3. Code for Equations *(4.2)* for $(i,n) \in \{2,\dots,i^*-1\} \times \{1,\dots,N-1\}$ . . . 60
4.3.4. Code for Equations *(4.2)* for $(i,n) \in \{i^*+1,\dots,I-1\} \times \{1,\dots,N-1\}$ 61
4.3.5. Code for Equations *(4.3)* . . . 62
4.3.6. Code for Equations *(4.5)* . . . 62

4.3.7. Code for Equations *(4.6)* 63

5. Iterative method **65**
5.1. Introduction 66
5.2. Preliminaries 66
5.3. Description of the iterative algorithm 67

6. Direct method versus Iterative method **71**
6.1. Prefatory comments 72
6.2. Test #1 72
6.2.1. Results obtained with the iterative method 73
6.2.2. Results obtained with the direct method 75
6.3. Test #2 76
6.3.1. Results obtained with the iterative method 77
6.3.2. Results obtained with the direct method 77
6.4. Some additional comments 78

7. Azimuthal variable dependent problem **79**
7.1. Introduction 80
7.2. Hypotheses 81
7.3. Fourier technique 84
7.4. Scheme for the θ-independent problem *(7.35)–(7.37)* 86
7.4.1. Description of the core scheme 88
7.4.2. Numerical experiments with the core scheme 90
7.5. Numerical results for the full problem 92
7.5.1. Test #1 93
7.5.2. Test #2 94

Conclusions and future work **97**

Resumen en castellano **99**

Bibliography **109**

Introduction

The goal of this thesis is to present direct and iterative methods based on a new finite difference scheme for computing the function

$$\psi : (\mu, z, \theta) \in Q = [-1, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}] \times [0, 2\pi) \to \psi(\mu, z, \theta) \subset \mathbb{R}, \ Q \subset \mathbb{R}^3$$

understanding that $Z_{\text{ini}}, Z_{\text{fin}} \in \mathbb{R}, Z_{\text{ini}} < Z_{\text{fin}}$ and that ψ is the solution of the problem defined by the degenerate parabolic PDE (notice that $\mu = 0$ eliminates $\frac{\partial \psi}{\partial z}$)

$$\mu \frac{\partial \psi}{\partial z} + \alpha\psi - \sigma \left\{ \frac{\partial}{\partial \mu} \left[(1 - \mu^2) \frac{\partial \psi}{\partial \mu} \right] + \frac{1}{1 - \mu^2} \frac{\partial^2 \psi}{\partial \theta^2} \right\} = W \text{ for } (\mu, z, \theta) \in Q$$

with the conditions

$$\begin{aligned}
&\psi_{|\{\mu \in (0,1],\, z = Z_{\text{ini}}\}} = f, \text{ with } f = f(\mu, \theta) \text{ given,} \\
&\psi_{|\{\mu \in [-1,0),\, z = Z_{\text{fin}}\}} = g, \text{ with } g = g(\mu, \theta) \text{ given,} \\
&\psi_{|\{\theta = 0\}} = \psi_{|\{\theta = 2\pi\}}, \left(\frac{\partial \psi}{\partial \theta} \right)_{|\{\theta = 0\}} = \left(\frac{\partial \psi}{\partial \theta} \right)_{|\{\theta = 2\pi\}}.
\end{aligned}$$

Function ψ is representing the angular flux density of charged particles, for example electrons. Variable z stands for 1D space, μ for the cosine of the polar angle and θ for the azimuthal angle. More details on numerical resolution for a more slightly general case than given below will be studied in Chapter 7.

From this Chapter on we will investigate azimuthal angle independent problem, that is the case $\psi = \psi(\mu, z)$, which reads as follows:

$$\mu \frac{\partial \psi}{\partial z} + \alpha\psi - \sigma \frac{\partial}{\partial \mu} \left[(1 - \mu^2) \frac{\partial \psi}{\partial \mu} \right] = W, \text{ for } (\mu, z) \in Q, \tag{1}$$

$$\psi(\mu, Z_{\text{ini}}) = f(\mu), \text{ for } \mu \in (0, 1], \tag{2}$$

$$\psi(\mu, Z_{\text{fin}}) = g(\mu), \text{ for } \mu \in [-1, 0). \tag{3}$$

The equation *(1)* is known as the degenerate parabolic or forward-backward equation and conditions *(2)*, *(3)* form boundary conditions.

Even for constant data, the solution ψ may be non-differentiable. In order to cover a larger class of admissible solutions, we shall assume, in a first stage, that ψ is continuous on Q. Speaking in classical terms such as "continuity" or "differentiability", no regularity result occurs in the mathematical literature; however, continuity represents well what is seen in all the available numerical results and that is precisely the reason why continuity has marked our mental paradigm of the exact solution for designing the numerical method.

Later on we shall see that the exact solution of the problem α constant and $\sigma \equiv W \equiv 0$ fails to be continuous if f or g is a nonzero constant, but when $\sigma > 0$ the diffusivity helps to regularize the solution at least up to continuity.

Let us fix now our working hypotheses.

In Equation *(1)*, $\alpha \geq 0, \sigma > 0$ and W are given functions of $(\mu, z) \in Q$, while f and g in Equations *(2)* and *(3)* are given functions of μ in $(0, 1]$ and $[-1, 0)$, respectively.

Inasmuch as Equations *(2)* and *(3)* imply that

$$f \in \mathrm{C}\big((0,1]\big), g \in \mathrm{C}\big([-1,0)\big), \text{ and} \tag{4}$$

$$\text{both } \lim_{\mu\downarrow 0} f(\mu) \text{ and } \lim_{\mu\uparrow 0} g(\mu) \text{ must exist in } \mathbb{R} \tag{5}$$

so that a continuous solution ψ exist, it will be assumed throughout this thesis that conditions *(4)* and *(5)* are fulfilled, and the notations $f(0)$ and $g(0)$ will be used with their obvious meanings. This amounts to say that f and g are supposed to be continuous in $[0,1]$ and $[-1,0]$, respectively.

Problem *(1)–(3)* is often found with $W \equiv 0$, but considering non-zero sources W is essential in this work as it allows building problems with known regular exact solutions and hence checking experimentally the order of convergence of the numerical scheme. Afterwards, the knowledge of the order can be employed to gain insight into the regularity of the the exact solution when no *a priori* knowledge about regularity exists. Moreover, the transient problem turns into a collection of steady problems with non-zero sources after time dicretization.

It is noteworthy that, despite the presence of the diffusive term

$$\sigma \frac{\partial}{\partial \mu}\Big[(1-\mu^2)\frac{\partial \psi}{\partial \mu}\Big],$$

no conditions at $\mu \in \{-1,1\}$ are needed. This fact has been already observed for similar problems. Beals considers in [3] the problem with $\alpha \equiv 0$, and says that the degeneracy of the diffusion operator in the μ-variable accounts for the lack of boundary conditions at $|\mu| = 1$. More recently, Epstein and Mazzeo, in their monograph [11], also address the question of the absence of boundary conditions for a class of partial differential equations that arise in population genetics and mathematical finance, and that resemble Equation *(1)* in that the diffusivity is a second degree polynomial with roots at the two boundary points. Ultimately, what mathematical analysis says is that the solution of the problem *(1)–(3)* must be looked for in an adequate functional space that guarantees existence and uniqueness without the need of any imposed values at $|\mu| = 1$; these values come out as a by-product of the resolution process, exactly in the same way as values at interior points do.

In what regards this question, we simply want to add that formal integration of $\big(D(\mu)u'(\mu)\big)' = 0$ in (a,b) with prescribed boundary values $u(a)$ and $u(b)$ provides one with

$$u(\mu) = u(a) \tag{6}$$

if $u(a) = u(b)$ (obviously, constant functions are solutions) and with

$$u(\mu) = u(a) + \big(u(b) - u(a)\big)\left(\int_a^b \frac{1}{D(s)}ds\right)^{-1} \int_a^\mu \frac{1}{D(s)}ds \tag{7}$$

if $u(a) \neq u(b)$. But Equation *(7)* is representing a solution only if $\frac{1}{D}$ is integrable on (a,b) and if the derivative of $\int_a^\mu \frac{1}{D(s)}ds$ is $\frac{1}{D(\mu)}$ for every $\mu \in (a,b)$; this setting covers indeed many situations, but it does not cover the case interest here, which is $D(\mu) = 1-\mu^2, (a,b) = (-1,1)$. In fact, the only solutions of $\big((1-\mu^2)u'(\mu)\big)' = 0$ in $(-1,1)$ that have finite limits at -1 and 1 are constant functions.[(a)] For this ODE, we could impose either one, and only one, of the

[(a)] Consider $\big((1-\mu^2)u'(\mu)\big)' = 0$ in $(-1,1)$. Then $(1-\mu^2)u'(\mu) = C_1$, $u'(\mu) = \frac{C_1}{1-\mu^2}$, $u(\mu) = C_2 + \frac{C_1}{2}\ln|\frac{1+\mu}{1-\mu}|$. Finally, $\lim_{\mu\downarrow -1}\ln|\frac{1+\mu}{1-\mu}| = -\infty$ and $\lim_{\mu\uparrow 1}\ln|\frac{1+\mu}{1-\mu}| = \infty$.

boundary values (in this case the solution would be constant), or two values at distinct interior points $\mu_1, \mu_2 \in (-1, 1)$.

In accordance with the discussion above, it might be erroneous to believe that prescribed values at the border points are not needed because the diffusivity at these points is zero; to take an example, the problem $\left(\sqrt{1-\mu^2}u'(\mu)\right)' = 0$ in $(-1, 1)$ admits arbitrary boundary conditions at $\{-1, 1\}$.[(b)] In other words, the lack of boundary conditions in the problem *(1)–(3)* seems to be related to the rate of convergence of the diffusivity towards zero when $|\mu| \to 1$, rather than to the mere fact that the diffusivity is null at these points.

The reader is referred to [4], [6], [12], [22], [23] and [29] for further theoretical considerations about this problem. In [24] there is a proof of existence and uniqueness for the problem *(1)–(3)* when α and σ are constant; the same reference shows that the condition $\alpha \geq 0$ is not necessary for the problem to have a unique solution, but we shall maintain it in order to preserve the physical "absorbing" meaning. In [29] there is beautiful proof of uniqueness when σ is constant and $\alpha \equiv W \equiv 0$.

Theoretical comments on this problem have an extraordinary importance, but they are not in the realm of this work, and we will not go back to them in what follows.

[(b)] Consider $\left(\sqrt{1-\mu^2}u'(\mu)\right)' = 0$ in $(-1, 1)$. Then $\sqrt{1-\mu^2}u'(\mu) = C_1$, $u'(\mu) = \frac{C_1}{\sqrt{1-\mu^2}}$, $u(\mu) = C_2 + C_1 \arcsin(\mu)$. Finally, $\arcsin(-1) = -\frac{\pi}{2}$ and $\arcsin(1) = \frac{\pi}{2}$.

Chapter 1

Derivation of 1D FPE from the general 3D FPE.

Contents

1.1. Introduction **6**

1.2. Physical domain **9**

1.3. Fourier techniques **9**

1.4. Simplifications **12**

1.4.1. First and second simplifications 12

1.4.2. Third and fourth simplifications 13

1.1. Introduction

The generic name "Fokker-Planck equation" actually houses a set of kinetic equations which are different from each other and which model different phenomena, from fluid flow to the movement of flocks of birds.

Therefore, this Section is not only to talk about the physical meaning, but also to specify that the FPE we are thinking of is the one modelling transport of charged particles like electrons. This equation comes from the field of nuclear engineering and reads as follows:

$$\frac{1}{c}\frac{\partial\psi}{\partial t} + \boldsymbol{\omega}\cdot\nabla\psi + \alpha\psi =$$

$$\sigma\left\{\frac{\partial}{\partial\mu}\left[(1-\mu^2)\frac{\partial\psi}{\partial\mu}\right] + \frac{1}{1-\mu^2}\frac{\partial^2\psi}{\partial\theta^2}\right\} + \frac{\partial\big(S_M\psi\big)}{\partial\epsilon} + W, \tag{1.1}$$

where

- $\psi = \psi(\boldsymbol{x}, t, \mu, \theta, \epsilon)$ is the *angular flux density* of particles.
- $\boldsymbol{x} = (x_1, x_2, x_3)$ is a point in some set $\Omega \subset \mathbb{R}^3$, being Ω the physical spatial domain.
- $t \in [T_{\text{ini}}, T_{\text{fin}}]$ stands for time.
- c is the speed of particles in the medium.
- $\boldsymbol{\omega} \in S^2$ stands for the direction of particle propagation. One has

$$\boldsymbol{\omega} = \boldsymbol{\omega}(\varphi, \theta) = (\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi), \tag{1.2}$$

  with $\varphi \in [0, \pi]$ the polar angle and $\theta \in [0, 2\pi)$ the azimuthal angle, in the standard spherical coordinate system centered at x. Here, S^2 is the unit sphere in $\mathbb{R}^3$.
- ∇ stands for the gradient with respect to the three spatial variables. Hence, $\big(\nabla\psi\big)_i = \frac{\partial\psi}{\partial x_i}$ for $i \in \{1, 2, 3\}$.
- $\mu = \omega_3 = \cos\varphi \in [-1, 1]$. Notice that μ uniquely determines the polar angle φ.
- $\epsilon \in (0, \infty)$ is the particle energy.

In their maximal generality, the data functions α, σ and W can depend on $(\boldsymbol{x}, t, \mu, \theta, \epsilon)$ when they appear in Equation *(1.1)*.

Terms accompanied by functions $\alpha \geq 0$ and $\sigma > 0$ account, respectively, for absorption and scattering phenomena, and $\sigma = \frac{\Sigma_{\text{tr}}}{2}$, where Σ_{tr} is the *momentum transfer* or *transport-corrected scattering cross section*.

Function $S_M = S_M(\boldsymbol{x}, \epsilon) > 0$ is called the *stopping power*, and represents an average energy loss per unit path length. It is given by the Bethe formula, including or not the Barkas and Bloch corrections on the basis of the demanded accuracy (see [34]). In reference [2],

for instance, one can find a clear explanation of the meaning of the stopping power when particles are electrons: "inelastic collisions occur so frequently that, as an approximation, electrons can be considered to undergo a continuous slowing down, with a fixed energy loss per unit path length travelled. This quantity is referred to as the stopping power and is well known both experimentally and theoretically". By the way, the same sentence explains why operator $\frac{\partial(S_M\cdot)}{\partial\epsilon}$ is known as the *continuous-slowing-down operator*.

Notations may differ from one to another reference, but to understand Equation *(1.1)* it suffices to say that W is a known volumetric source (wherever it is positive) or sink (wherever it is negative) placed in the interior of the domain, and to refer the reader to the bibliography for the standard meaning of all the other terms in each specific application. Apart from references [13], [19] and [21] which are at the origin of the present work, the reader's vision will be enriched by studying some review from the nuclear engineering community such as reference [24].

Suppose that Ω is non-empty, open, bounded and convex, having boundary $\partial\Omega$ (which under the present hypotheses is necessarily Lipschitz (see [16])), and let $\boldsymbol{n}(\boldsymbol{x})$ be the outward unit normal at $\boldsymbol{x}\in\partial\Omega$. Then, a possible set of conditions that closes Equation *(1.1)* is the following:

$$\psi_{|\{t=T_{\text{ini}}\}} \equiv \widehat{\psi} \text{ with } \widehat{\psi}=\widehat{\psi}(\boldsymbol{x},\mu,\theta,\epsilon) \text{ given} \tag{1.3}$$

$$\psi_{|\{(\boldsymbol{x},\boldsymbol{\omega})\in\partial\Omega\times S^2:\,\boldsymbol{\omega}\cdot\boldsymbol{n}(\boldsymbol{x})<0\}} \equiv \mathscr{F} \text{ with } \mathscr{F}=\mathscr{F}(\boldsymbol{x},t,\mu,\theta,\epsilon) \text{ given,} \tag{1.4}$$

$$\psi_{|\{(\theta=0\}} \equiv \psi_{|\{\theta=2\pi\}}, \quad \left(\frac{\partial\psi}{\partial\theta}\right)_{|\{(\theta=0\}} \equiv \left(\frac{\partial\psi}{\partial\theta}\right)_{|\{\theta=2\pi\}}, \tag{1.5}$$

$$\psi_{|\{\epsilon=\infty\}} \equiv 0. \tag{1.6}$$

When c in the FPE *(1.1)* is the speed of light or close to, the variation of ψ with time is not, in a wide range of applications, so fast as the non-steady term $\frac{1}{c}\frac{\partial\psi}{\partial t}$ must be taken into account, and frequently the steady FPE, rather than the transient one, is solved. Naturally, the steady case obliges to eliminate t as independent variable in all functions above, and it does not require the initial condition *(1.3)*.

Existence and uniqueness of solution for a very similar problem to the one defined by Equations *(1.1)*–*(1.6)*, but in the steady case, has been proved in reference [20].

The initial condition *(1.3)* is linked to the transient term $\frac{\partial\psi}{\partial t}$, the boundary condition *(1.4)* to the transport term $\boldsymbol{\omega}\cdot\nabla\psi$, the periodicity conditions *(1.5)* to $\frac{\partial^2\psi}{\partial\theta^2}$, and the condition at infinite energy *(1.6)* to $\frac{\partial(S_M\psi)}{\partial\epsilon}$. As we are now placed in the physical context, we do not like to call periodic boundary conditions to conditions *(1.5)*, despite they can be rightly called in this way from a mathematical viewpoint, as they do not refer at all to the physical boundary. A "final" condition for energy is needed, and not an initial one, due to the positivity of σ and S_M (think of the heat equation being ϵ the time variable).

The initial condition *(1.3)* simply states that the angular flux density at the initial time is known.

Through condition *(1.4)* the incoming particle flux is imposed. One could use another type of condition, for instance (see [23]) an $\mathscr{F}$ depending on the function ψ itself to account

for an eventual reflected flux, but it is not possible to get rid of the set $\{(\boldsymbol{x},\boldsymbol{\omega}) \in \partial\Omega \times S^2 : \boldsymbol{\omega}\cdot\boldsymbol{n}(\boldsymbol{x}) < 0\}$ over which the condition is posed.

The interpretation of conditions *(1.5)* is clear from the geometric meaning of azimuthal angle θ, as $\theta = 0$ and $\theta = 2\pi$ are defining exactly the same direction $\boldsymbol{\omega}$ once the polar angle φ is fixed.

Lastly, condition *(1.6)* simply states that there are not particles having infinite energy.

While several possibilities exist for condition *(1.4)*, conditions *(1.3)*, *(1.5)* and *(1.6)* are needed in the way they have been stated here.

Notice that the term $\frac{\partial}{\partial\mu}\left[(1-\mu^2)\frac{\partial\psi}{\partial\mu}\right]$ does not provide any condition at $|\mu| = 1$. A somewhat casual explanation on this fact is as follows: on the one hand, periodicity has no sense as $\mu = -1$ and $\mu = 1$ are never defining the same direction; on the other hand, other type of condition, like Dirichlet one, violates the Physics, because $|\mu| = 1$ is not defining a physical boundary; in other words, the flux is determined by the incoming flux through the (physical) boundary $\partial\Omega$, and it cannot be imposed at interior points of the physical domain for any value of μ. It is a nice property of the mathematical model that there are also analytical reasons for these conditions not to be needed.

Operator

$$\frac{\partial}{\partial\mu}\left[(1-\mu^2)\frac{\partial\cdot}{\partial\mu}\right] + \frac{1}{1-\mu^2}\frac{\partial^2\cdot}{\partial\theta^2}.$$

is known as the *continuous-scattering operator*. Since this operator is precisely the *spherical Laplacian* (also known as the *Laplace-Beltrami operator*), the FPE is imposing, on the $\boldsymbol{\omega}$ variable, simple diffusion on S^2, with a coefficient of diffusivity given by the function σ. The FPE thus connects with the interesting field of partial differential equations on surfaces, called more succinctly *surface PDEs*.

The importance of Equation *(1.1)* lies in that its solution is an approximation of the solution to the Boltzmann transport equation for particles that suffer highly forward-peaked scattering and small energy losses, and it can serve as a model for diverse phenomena: in [13] and [19], the authors consider the case $\alpha \equiv W \equiv 0$ in order to model electron transport through human body, which has applications in external radiotherapy; in [21], the authors take $W \equiv 0$ and eliminate the energy dependence, hence removing the term $\frac{\partial\left(S_M\psi\right)}{\partial\epsilon}$, in order to study light propagation in biological tissues, which has applications in tomographic imaging (see [23]). All papers which have been cited within this paragraph use the steady FPE.

Both FPE and BTE are equations for the angular flux density of particles and they state the balance between gains and losses for $\omega\cdot\nabla\psi$, the directional derivative of ψ along each direction of propagation ω.

1.2. Physical domain

Let $\Omega \subset \mathbb{R}$ be the spatial domain. When ψ depends on position $\boldsymbol{x} = (x_1, x_2, x_3) \in \Omega$ only through $x_3 = z$, then $\boldsymbol{\omega} \cdot \nabla\psi = \omega_3 \frac{\partial \psi}{\partial z}$, which corresponds to the first term of Equation *(1)* if we understand that $\mu = \omega_3$. Under previous assumption, we can say that the problem is posed in the 1D slab if, additionally the spatial domain has got the form $\Omega = \Omega^* \times (Z_{\text{ini}}, Z_{\text{fin}})$, with $\Omega^* \subset \mathbb{R}^2$ (see Figure 1.1). The idea is that spatial variables can be reduced from 3 to

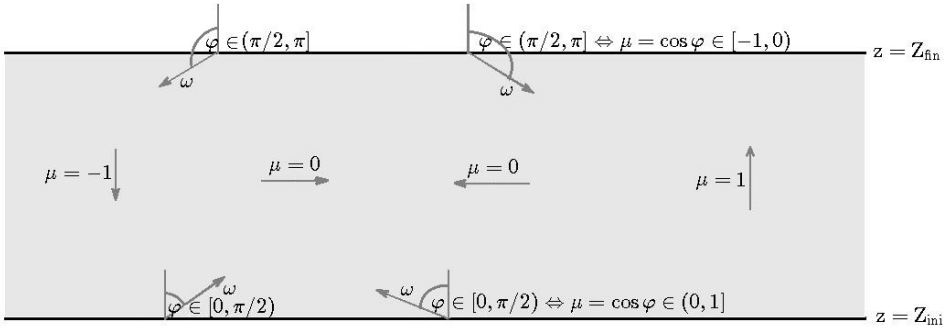


Figure 1.1: (Section of) the one-dimensional slab. One must take into account that this is a special kind of three-dimensional domain.

1 because of two reasons: ψ does not vary either with x_1 or with x_2, and the domain is a collection of copies of the 1D domain $(Z_{\text{ini}}, Z_{\text{fin}})$.

1.3. Fourier techniques

Periodicity conditions *(1.5)* invite to use Fourier techniques to reduce the problem *(1.1)–(1.6)* to a collection of θ-independent problems.

Suppose that the angular flux density can be expressed as a Fourier series

$$\psi(\boldsymbol{x}, t, \mu, \theta, \epsilon) = \sum_{k=-\infty}^{\infty} \psi_k(\boldsymbol{x}, t, \mu, \epsilon)\, \mathrm{e}^{ik\theta}, \tag{1.7}$$

the coefficients of which are given by

$$\psi_k(\boldsymbol{x}, t, \mu, \epsilon) = \frac{1}{2\pi} \int_0^{2\pi} \psi(\boldsymbol{x}, t, \mu, \theta, \epsilon)\, \mathrm{e}^{-ik\theta} d\theta. \tag{1.8}$$

In case that the series in Equation *(1.7)* converges rapidly, one can get ψ if only a few coefficients ψ_k are known. That is why we will investigate now which transport equation is satisfied by the Fourier coefficients ψ_k.

Given a function G, G_k will denote its k^{th} Fourier coefficient with respect to the variable θ.

The first thing to do, following Equation *(1.8)*, is to multiply each term in Equation *(1.1)* by $\mathrm{e}^{-ik\theta}$, then integrate from $\theta =$ to $\theta = 2\pi$, and finally divide by 2π. We need to assume that neither α nor σ depends on θ. The results are the following:

$$\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{c}\frac{\partial\psi}{\partial t}\,\mathrm{e}^{-ik\theta}d\theta = \frac{1}{c}\frac{\partial\psi_k}{\partial t}, \tag{1.9}$$

$$\frac{1}{2\pi}\int_0^{2\pi}(\boldsymbol{\omega}\cdot\nabla\psi)\,\mathrm{e}^{-ik\theta}d\theta =$$

$$\frac{\sqrt{1-\mu^2}}{2}\left\{\frac{\partial(\psi_{k+1}+\psi_{k-1})}{\partial x_1} + i\frac{\partial(\psi_{k+1}-\psi_{k-1})}{\partial x_2}\right\} + \mu\frac{\partial\psi_k}{\partial x_3}, \tag{1.10}$$

$$\frac{1}{2\pi}\int_0^{2\pi}\alpha\psi\,\mathrm{e}^{-ik\theta}d\theta = \alpha\psi_k, \tag{1.11}$$

$$\frac{1}{2\pi}\int_0^{2\pi}\sigma\frac{\partial}{\partial\mu}\left\{(1-\mu^2)\frac{\partial\psi}{\partial\mu}\right\}\mathrm{e}^{-ik\theta}d\theta =$$

$$\sigma\frac{\partial}{\partial\mu}\left[(1-\mu^2)\frac{\partial\psi_k}{\partial\mu}\right], \tag{1.12}$$

$$\frac{1}{2\pi}\int_0^{2\pi}\frac{\sigma}{1-\mu^2}\frac{\partial^2\psi}{\partial\theta^2}\,\mathrm{e}^{-ik\theta}d\theta = -\frac{\sigma k^2}{1-\mu^2}\psi_k, \tag{1.13}$$

$$\frac{1}{2\pi}\int_0^{2\pi}\frac{\partial(S_M\psi)}{\partial\epsilon}\,\mathrm{e}^{-ik\theta}d\theta = \frac{\partial(S_M\psi_k)}{\partial\epsilon}, \tag{1.14}$$

$$\frac{1}{2\pi}\int_0^{2\pi}W\,\mathrm{e}^{-ik\theta}d\theta = W_k. \tag{1.15}$$

Equation *(1.10)* is the most difficult to obtain, and that is why we take some lines to explain it in detail. Firstly, one must take into account Equation *(1.2)*. Then

$$\frac{1}{2\pi}\int_0^{2\pi}(\boldsymbol{\omega}\cdot\nabla\psi)\,\mathrm{e}^{-ik\theta}d\theta =$$

$$\frac{1}{2\pi}\left\{\int_0^{2\pi}\sin\varphi\cos\theta\frac{\partial\psi}{\partial x_1}\,\mathrm{e}^{-ik\theta}d\theta + \int_0^{2\pi}\sin\varphi\sin\theta\frac{\partial\psi}{\partial x_2}\,\mathrm{e}^{-ik\theta}d\theta + \right. \tag{1.16}$$

$$\left.\int_0^{2\pi}\cos\varphi\frac{\partial\psi}{\partial x_3}\,\mathrm{e}^{-ik\theta}d\theta\right\}.$$

Since $\mu = \cos\varphi$, the third summand above provides one with the term $\mu\frac{\partial\psi_k}{\partial x_3}$ in Equation *(1.10)*; also $\sin\varphi = \sqrt{1-\mu^2}$ accounts for the presence of this factor in the same equation. Finally, the key point in obtaining Equation *(1.10)* is that

$$\big[(\cos\theta)\psi\big]_k = \frac{1}{2}(\psi_{k+1}+\psi_{k-1}), \tag{1.17}$$

$$\big[(\sin\theta)\psi\big]_k = \frac{i}{2}(\psi_{k+1}-\psi_{k-1}). \tag{1.18}$$

Equalities *(1.17)* and *(1.18)* above can be checked by direct calculation or, alternatively, by using computations on products of Fourier series in [31, Section 5, par. 6].

Equation *(1.13)* is the result of integrating twice by parts and taking into account the periodic conditions *(1.5)*. Indeed,

$$\int_0^{2\pi} \frac{\partial^2 \psi}{\partial \theta^2} \mathrm{e}^{-ik\theta} d\theta = \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=2\pi\}} - \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=0\}} + ik(\psi_{|\{\theta=2\pi\}} - \psi_{|\{\theta=0\}}) - k^2 \int_0^{2\pi} \psi \, \mathrm{e}^{-ik\theta} d\theta = -k^2 \int_0^{2\pi} \psi \, \mathrm{e}^{-ik\theta} d\theta. \tag{1.19}$$

The rest of Equations *(1.9)*–*(1.15)* are trivial. Now, adding all of them, one arrives at

$$\frac{1}{c}\frac{\partial \psi_k}{\partial t} + \frac{\sqrt{1-\mu^2}}{2}\left\{\frac{\partial(\psi_{k+1}+\psi_{k-1})}{\partial x_1} + i\frac{\partial(\psi_{k+1}-\psi_{k-1})}{\partial x_2}\right\} + \mu\frac{\partial \psi_k}{\partial x_3} + \left(\alpha + \frac{\sigma k^2}{1-\mu^2}\right)\psi_k = \sigma\frac{\partial}{\partial \mu}\left[(1-\mu^2)\frac{\partial \psi_k}{\partial \mu}\right] + \frac{\partial(S_M\psi_k)}{\partial \epsilon} + W_k. \tag{1.20}$$

The problem with Equation *(1.20)* is that it needs a closure for ψ_{k-1} and ψ_{k+1}, which are unknowns appearing in what we wanted to be a single equation for ψ_k. Even if we thought of a system with several unknowns $\psi_{k-m}, \ldots, \psi_{k+m}$ with $m \geq 2$, we would need again a closure for ψ_{k-m} and ψ_{k+m}.

It is also clear that we would obtain the desired single equation for ψ_k in case that ψ does not depend either on x_1 or on x_2.

The conclusion is that θ-dependence can be eliminated by means of Fourier techniques if ψ depends on $\boldsymbol{x}$ only through x_3, and besides, neither α nor σ depends on θ. Under these hypotheses, one has, for each Fourier coefficient ψ_k

$$\frac{1}{c}\frac{\partial \psi_k}{\partial t} + \mu\frac{\partial \psi_k}{\partial x_3} + \left(\alpha + \frac{\sigma k^2}{1-\mu^2}\right)\psi_k = \sigma\frac{\partial}{\partial \mu}\left[(1-\mu^2)\frac{\partial \psi_k}{\partial \mu}\right] + \frac{\partial(S_M\psi_k)}{\partial \epsilon} + W_k, \tag{1.21}$$

closed with the following conditions:

$$(\psi_k)_{|\{t=T_{\mathrm{ini}}\}} \equiv \widehat{\psi}_k, \tag{1.22}$$

$$(\psi_k)_{|\{(\boldsymbol{x},\boldsymbol{\omega})\in\partial\Omega\times S^2:\, \boldsymbol{\omega}\cdot\boldsymbol{n}(\boldsymbol{x})<0\}} \equiv \mathscr{F}_k, \tag{1.23}$$

$$(\psi_k)_{|\{\epsilon=\infty\}} \equiv 0. \tag{1.24}$$

1.4. Simplifications

Some simplifications detailed below will turn the seven-dimensional problem *(1.1)–(1.6)* into the two-dimensional problem *(1)–(3)*.

1.4.1. First and second simplifications

Assume firstly that the problem is, or can be treated as, steady.

Assume secondly as in the previous Section that the problem depends on $\boldsymbol{x}$ only through $x_3 = z$. This is both a geometrical and a functional assumption: on the one hand, to the already required properties for the domain we add that it can be written as $\Omega = \Omega^* \times (Z_{\text{ini}}, Z_{\text{fin}})$, with $\Omega^* \subset \mathbb{R}^2$, and, on the other hand, we assume that $\psi = \psi(z, \mu, \theta, \epsilon)$. In such a case, $\boldsymbol{\omega} \cdot \nabla\psi = \mu \frac{\partial \psi}{\partial z}$ and consequently Equation *(1.1)* gets simplified into

$$\mu \frac{\partial \psi}{\partial z} + \alpha\psi = \sigma \left\{ \frac{\partial}{\partial \mu} \left[(1-\mu^2) \frac{\partial \psi}{\partial \mu} \right] + \frac{1}{1-\mu^2} \frac{\partial^2 \psi}{\partial \theta^2} \right\} + \frac{\partial (S_M \psi)}{\partial \epsilon} + W. \tag{1.25}$$

In Equation *(1.25)*, functions α, σ and W depend on $(z, \mu, \theta, \epsilon)$.

Boundary conditions are needed only at $z = Z_{\text{ini}}$ (lower face of Ω) and $z = Z_{\text{fin}}$ (upper face of Ω) because, from a mathematical perspective, Equation *(1.25)* is one dimensional in space. As the solution ψ does not depend on x_1 and x_2 and the spatial domain is a collection of copies of $(Z_{\text{ini}}, Z_{\text{fin}})$, the knowledge of ψ on one single copy determines the solution on the whole Ω (see Figure 1.1).

Now notice that for boundary points $\boldsymbol{x}$ with $x_3 = z = Z_{\text{ini}}$ we have $\boldsymbol{n}(\boldsymbol{x}) = (0, 0, -1)$, which implies $\boldsymbol{\omega} \cdot \boldsymbol{n}(\boldsymbol{x}) = -\omega_3 = -\mu$. Analogously, $\boldsymbol{\omega} \cdot \boldsymbol{n}(\boldsymbol{x}) = \omega_3 = \mu$ when $z = Z_{\text{fin}}$. Therefore, the boundary conditions for Equation *(1.25)* are, instead of *(1.4)*, the following ones:

$$\psi_{|\{z=Z_{\text{ini}},\ \mu\in(0,1]\}} = \mathscr{F}_{\text{ini}}, \text{ with } \mathscr{F}_{\text{ini}} = \mathscr{F}_{\text{ini}}(\mu, \theta, \epsilon) \text{ given}, \tag{1.26}$$

$$\psi_{|\{z=Z_{\text{fin}},\ \mu\in[-1,0)\}} = \mathscr{F}_{\text{fin}}, \text{ with } \mathscr{F}_{\text{fin}} = \mathscr{F}_{\text{fin}}(\mu, \theta, \epsilon) \text{ given}, \tag{1.27}$$

and the problem gets closed by adding the periodicity conditions *(1.5)* and the condition at infinite energy *(1.6)*, which are repeated here so that the full problem can be cited in the sequel by means of a set of consecutive equations:

$$\psi_{|\{\theta=0\}} \equiv \psi_{|\{\theta=2\pi\}}, \quad \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=0\}} \equiv \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=2\pi\}}, \tag{1.28}$$

$$\psi_{|\{\epsilon=\infty\}} \equiv 0. \tag{1.29}$$

The set $\Omega^* \subset \mathbb{R}^2$ plays really no role in the resolution process, which motivates some authors to think of the domain Ω as an infinite slab limited by the parallel planes $z = Z_{\text{ini}}$ and $z = Z_{\text{fin}}$: $\Omega = \mathbb{R}^2 \times \big(Z_{\text{ini}}, Z_{\text{fin}}\big)$. In this setting one says that the problem is posed in *the one-dimensional slab*.

We remark that there are other simplifications, for instance when certain symmetries can be assumed in cylindrical or spherical geometry, that, in the case explained above, result in a problem which is one-dimensional in space. We do not pursue here these approximations.

1.4.2. Third and fourth simplifications

If moreover θ-dependence is neglected (third simplification), i.e. $\psi = \psi(z, \mu, \epsilon)$, then Equation *(1.25)* turns into

$$\mu \frac{\partial \psi}{\partial z} + \alpha \psi = \sigma \frac{\partial}{\partial \mu} \left[(1 - \mu^2) \frac{\partial \psi}{\partial \mu} \right] + \frac{\partial (S_M \psi)}{\partial \epsilon} + W. \tag{1.30}$$

This equation is closed with conditions

$$\psi_{|\{z=Z_{\text{ini}},\ \mu \in (0,1]\}} = \mathscr{F}_{\text{ini}}, \text{ with } \mathscr{F}_{\text{ini}} = \mathscr{F}_{\text{ini}}(\mu, \epsilon) \text{ given}, \tag{1.31}$$

$$\psi_{|\{z=Z_{\text{fin}},\ \mu \in [-1,0)\}} = \mathscr{F}_{\text{fin}}, \text{ with } \mathscr{F}_{\text{fin}} = \mathscr{F}_{\text{fin}}(\mu, \epsilon) \text{ given}, \tag{1.32}$$

$$\psi_{|\{\epsilon=\infty\}} \equiv 0, \tag{1.33}$$

which are conditions *(1.31)* and *(1.32)* free from θ-dependence, and condition *(1.33)*. When one reads that the steady FPE is being thought with *planar-geometry symmetry*, what is meant is exactly this problem, with possibly some variants of conditions *(1.31)*, *(1.32)*, as explained above. Planar-geometry symmetry is often assumed when the problem is posed in the one-dimensional slab.

Naturally, in Equation *(1.30)* the data functions α, σ and W are allowed depending on (z, μ, ϵ).

It is noteworthy to mention another way of understanding Equation *(1.30)* which is useful if we admit that α might be unbounded. Indeed, according to the discussion in Section 1.3, when both the absorption and scattering coefficients do not depend on θ, problem *(1.25)–(1.29)* can be reduced, via Fourier series, to a collection of problems of the type *(1.30)–(1.33)* (see also references [21] and [25]). The unboundedness of the new absorption coefficient, due to the presence of $1 - \mu^2$ in the denominator, is apparent when looking at Equation *(1.21)*.

The fourth and last simplification we consider is to eliminate energy dependence, which finally reduces the problem above to the problem *(1)-(3)*. This can mean that not only energy dependence is really being neglected, but also a prior energy discretization has been performed which splits the energy dependent problem into several energy independent problems.

Let us explain the last assertion in the paragraph above. If the energy spectrum is discretized via

$$\epsilon_1 = 0 < \epsilon_2 < \cdots < \epsilon_{P-1} < \epsilon_P,$$

with ϵ_P large enough to assume $\psi(\epsilon_P) \equiv 0$, then $\psi(\epsilon_{P-1}), \ldots, \psi(\epsilon_1)$ can be obtained, in this order, by integrating Equation *(1.30)* in the adequate energy interval or *energy group*. Let us show the details for $\psi(\epsilon_{P-1})$; in this case, Equation *(1.30)* must be integrated between ϵ_{P-1} and ϵ_P taking into account that

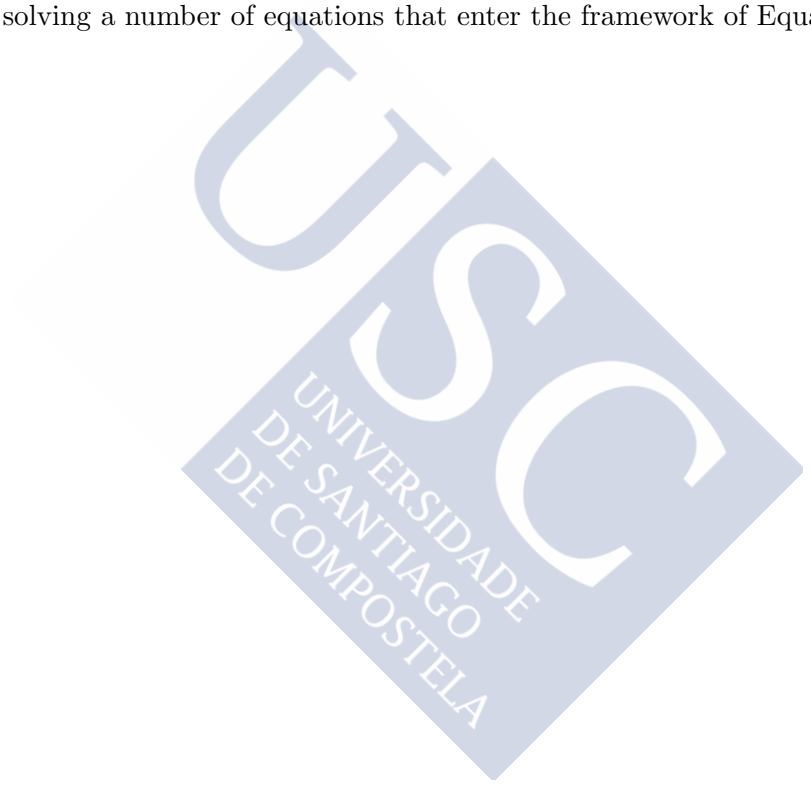
$$\int_{\epsilon_{P-1}}^{\epsilon_P} \frac{\partial (S_M \psi)}{\partial \epsilon} d\epsilon = -S_M(\epsilon_{P-1}) \psi(\epsilon_{P-1}),$$

which easily provides a new equation for $\psi(\epsilon_{P-1})$, of type *(1)*, if the integral $\int_{\epsilon_{P-1}}^{\epsilon_P} \psi(\epsilon)d\epsilon$ is approximated by trapezoidal quadrature:

$$\int_{\epsilon_{P-1}}^{\epsilon_P} \psi(\epsilon)d\epsilon \approx (\epsilon_P - \epsilon_{P-1})\frac{\psi(\epsilon_{P-1}) + \psi(\epsilon_P)}{2} = \frac{\epsilon_P - \epsilon_{P-1}}{2}\psi(\epsilon_{P-1}).$$

At this stage it is clear that the physical meaning of conditions *(2)* and *(3)* used in Introduction is not that of initial and final conditions, but rather boundary conditions imposing the incoming flux. For the purpose of designing the numerical method it is however useful to think of them as initial and final conditions, as will become clear later.

It is also clear that, by means of Fourier techniques and energy discretizations, the complete steady Fokker-Planck equation in the one-dimensional slab, i.e., Equation *(1.25)*, can be solved by solving a number of equations that enter the framework of Equation *(1)*.



Chapter 2

Analysis of a particular case

Contents

2.1. Trivial problem: case without diffusion 16

2.2. A particular example 16

2.2.1. Regularity analysis of function ψ 17

2.2.2. Continuity analysis 18

From this chapter on we change the order of the variables from (z, μ) to (μ, z) in order to keep similarities of the notations used in the references by which we have been motivated.

Let us call $Q_- = [-1, 0) \times [Z_{\text{ini}}, Z_{\text{fin}}]$, $Q_+ = (0, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}]$. We could say that we have a "final value problem" FVP on Q_- and "initial value problem" IVP on Q_+, but this assertion may be misleading, because these two problems are not independent except in the trivial situation $\sigma \equiv 0$, which *is not* our case. As it is illustrative, let us explain this point a little bit more.

2.1. Trivial problem: case without diffusion

If there is no diffusion ($\sigma \equiv 0$), then Equation *(1)* is ordinary if $\mu \neq 0$, and the problem can be split into two independent problems: one posed on Q_- and another one on Q_+. In detail:

- $\psi_{|Q_-}$ can be computed with the knowledge of the final datum $g(\mu)$ in condition *(3)* and
- $\psi_{|Q_+}$ can be computed with the knowledge of the initial datum $f(\mu)$ in condition *(2)*.

Moreover, the values of $\psi_{|\{\mu=0\}}$ are easily determined by taking $\mu = 0$ and $\sigma \equiv 0$ in Equation *(1)*:

$$\psi(0, z) = \frac{W(0, z)}{\alpha(0, z)}, \ \forall z \in [Z_{\text{ini}}, Z_{\text{fin}}]. \tag{2.1}$$

From Equation *(2.1)* it is inferred that conditions

$$f(0) = \frac{W(0, Z_{\text{ini}})}{\alpha(0, Z_{\text{ini}})} \ \text{ and } \ g(0) = \frac{W(0, Z_{\text{fin}})}{\alpha(0, Z_{\text{fin}})} \tag{2.2}$$

are necessary for ψ to be continuous, a limitation which does no longer hold if there is diffusion.

In this Subsection it has been assumed that $\alpha(0, z) > 0$ for all z. The reader can easily study the problem in case that $\alpha(0, z) = 0$ for some z.

2.2. A particular example

As an example, the function

$$\psi(\mu, z) = \begin{cases} g(\mu)\, \exp\left(\frac{\alpha(Z_{\text{fin}} - z)}{\mu}\right) & \text{if } (\mu, z) \in Q_-, \\ 0 & \text{if } (\mu, z) \in \{0\} \times [Z_{\text{ini}}, Z_{\text{fin}}], \\ f(\mu)\, \exp\left(-\frac{\alpha(z - Z_{\text{ini}})}{\mu}\right) & \text{if } (\mu, z) \in Q_+ \end{cases} \tag{2.3}$$

is the exact solution of problem *(1)–(3)* when α is a positive constant, $\sigma \equiv 0$ and $W \equiv 0$.

Since

$$\lim_{\mu\downarrow 0}\frac{\mathrm{e}^{-\frac{k}{\mu}}}{\mu^m}=\lim_{\mu\uparrow 0}\frac{\mathrm{e}^{\frac{k}{\mu}}}{\mu^m}=0$$

for any $k>0$ and any $m\in\mathbb{N}\cup\{0\}$,[(c)] the solution and its derivatives easily go to zero at interior points of Q when μ goes to zero, a fact that obliges f and g to be flat at $\mu=0$ to get a solution ψ with some chances of being regular at $(0,Z_{\mathrm{ini}})$ and $(0,Z_{\mathrm{fin}})$. This question is analysed in detail below.

2.2.1. Regularity analysis of function ψ

One says that $f\in \mathrm{C}^k\big([0,1]\big)$, with $k\in\mathbb{N}$, when f is k times differentiable on $[0,1]$ (from the right at $\mu=0$, from the left at $\mu=1$) and $f^{k)}$ is continuous on $[0,1]$. The assertion $g\in \mathrm{C}^k\big([-1,0]\big)$ must be understood in an analogous way.

Notice that $f\in \mathrm{C}^1\big([0,1]\big)$ if, and only if

1. $f\in \mathrm{C}\big([0,1]\big)$, and

2. f' exists and is continuous on $(0,1)$ and can be extended with continuity to 0 and 1.

To check the equivalence it is enough to pay attention to $\mu=0$. If $f\in \mathrm{C}^1\big([0,1]\big)$ then there exists

$$\lim_{\mu\downarrow 0} f'(\mu)=f'(0)\in\mathbb{R}.$$

Reciprocally, if $f\in \mathrm{C}\big([0,1]\big)\cap \mathrm{C}^1(0,1)$ and there exists $\lim_{\mu\downarrow 0} f'(\mu)\in\mathbb{R}$ then necessarily there exists

$$f'(0)=\lim_{\mu\downarrow 0}\frac{f(\mu)-f(0)}{\mu}=\lim_{\mu\downarrow 0} f'(\mu),$$

in virtue of the L'Hôpital's rule.

A function defined on $Q=[-1,1]\times[Z_{\mathrm{ini}},Z_{\mathrm{fin}}]$ is said to be of class 1 on Q, or to belong to $\mathrm{C}^1(Q)$, if

1. It is continuous on Q, and

2. Its partial derivatives with respect to μ and z exist and are continuous on Q° (the topological interior of Q) and can be extended with continuity up to the border of Q. We say for the sake of brevity that *its partial derivatives with respect to μ and z exist and are continuous on Q*, understanding that the values of the derivatives at the boundary points are the values of the appropriate limits.

Also, $\psi\in \mathrm{C}^k(Q)$, with $k\in\mathbb{N}$, $k\geq 2$, means that all its partial derivatives of order $k-1$ are of class 1 on Q.

[(c)] Of course, the limit is also zero if m is a negative integer, but this fact is not of interest here.

2.2.2. Continuity analysis

As

$$\psi_{|Q_+}(\mu, Z_{\text{ini}}) = f(\mu),\ \psi_{|Q_-}(\mu, Z_{\text{fin}}) = g(\mu),$$

and, for every $(\mu, z) \in Q$,

$$|\psi(\mu, z)| \leq \begin{cases} |g(\mu)| & \text{if } \mu < 0, \\ 0 & \text{if } \mu = 0, \\ |f(\mu)| & \text{if } \mu > 0, \end{cases}$$

function ψ is continuous on Q if, and only if, $f \in \mathrm{C}\big([0,1]\big)$, $g \in \mathrm{C}\big([-1,0]\big)$, and $f(0) = g(0) = 0$.

A finer analysis shows that continuity at the isolated point $(0, Z_{\text{ini}})$ holds with a less restrictive condition on g. Indeed, take

$$B = \{(\mu, z) \in Q :\ \| \ (\mu, z) - (0, Z_{\text{ini}}) \ \| < \delta\},$$

with $0 < \delta < \min\{1, Z_{\text{fin}} - Z_{\text{ini}}\}$ and notice that, for $(\mu, z) \in B$,

$$|\psi(\mu, z)| \leq \begin{cases} |g(\mu)| \exp\left(\frac{\alpha(Z_{\text{fin}} - Z_{\text{ini}} - \delta)}{\mu}\right) & \text{if } \mu < 0, \\ 0 & \text{if } \mu = 0, \\ |f(\mu)| & \text{if } \mu > 0. \end{cases}$$

Hence, it suffices that f is continuous at $\mu = 0,\ f(0) = 0$, and

$$\lim_{\mu \uparrow 0} \left\{ g(\mu) \exp\left(\frac{\alpha(Z_{\text{fin}} - Z_{\text{ini}} - \delta)}{\mu}\right)\right\} = 0$$

for ψ to be continuous at $(0, Z_{\text{ini}})$. For instance, we could have $g(\mu) = \frac{1}{\mu^m}$ with $m \in \mathbb{N}$.

Analogously, it suffices that g is continuous at $\mu = 0,\ g(0) = 0$, and

$$\lim_{\mu \downarrow 0} \left\{ f(\mu) \exp\left(\frac{-\alpha(Z_{\text{fin}} - Z_{\text{ini}} - \delta)}{\mu}\right)\right\} = 0$$

for ψ to be continuous at $(0, Z_{\text{fin}})$.

As soon as we want to maintain continuity at both points $(0, Z_{\text{ini}})$ and $(0, Z_{\text{fin}})$ simultaneously we need f and g to be continuous at $\mu = 0$ and $f(0) = g(0) = 0$.

Class 1 analysis.

- Notice firstly that $\frac{\partial \psi}{\partial z}$ exists and is continuous on Q if, and only if, $f \in \mathrm{C}\big([0,1]\big)$, $g \in \mathrm{C}\big([-1,0]\big)$, and $\lim_{\mu \downarrow 0} \frac{f(\mu)}{\mu} = \lim_{\mu \uparrow 0} \frac{g(\mu)}{\mu} = 0$. In fact,

$$\frac{\partial \psi}{\partial z}(\mu, z) = \begin{cases} -\frac{\alpha g(\mu)}{\mu} \exp\left(\frac{\alpha(Z_{\text{fin}} - z)}{\mu}\right) & \text{if } (\mu, z) \in Q_-, \\ 0 & \text{if } (\mu, z) \in \{0\} \times [Z_{\text{ini}}, Z_{\text{fin}}], \\ -\frac{\alpha f(\mu)}{\mu} \exp\left(-\frac{\alpha(z - Z_{\text{ini}})}{\mu}\right) & \text{if } (\mu, z) \in Q_+. \end{cases}$$

A detailed proof can be done by observing that

$$\Big(\frac{\partial\psi}{\partial z}\Big)_{|_{Q_+}}(\mu, Z_{\text{ini}}) = -\frac{\alpha f(\mu)}{\mu}, \quad \Big(\frac{\partial\psi}{\partial z}\Big)_{|_{Q_-}}(\mu, Z_{\text{fin}}) = -\frac{\alpha g(\mu)}{\mu},$$

and by using the inequalities

$$\Big|\frac{\partial\psi}{\partial z}(\mu, z)\Big| \leq \begin{cases} \alpha\Big|\frac{g(\mu)}{\mu}\Big| & \text{if } \mu < 0, \\ 0 & \text{if } \mu = 0, \\ \alpha\Big|\frac{f(\mu)}{\mu}\Big| & \text{if } \mu > 0, \end{cases}$$

valid for every $(\mu, z) \in Q$.

- On the other hand, $\frac{\partial\psi}{\partial\mu}$ exists and is continuous on Q if, and only if, $f \in \mathrm{C}^1\big([0,1]\big)$, $g \in \mathrm{C}^1\big([-1,0]\big)$, and $\lim_{\mu\downarrow 0}\frac{f(\mu)}{\mu^2} = \lim_{\mu\uparrow 0}\frac{g(\mu)}{\mu^2} = 0$. In fact,

$$\frac{\partial\psi}{\partial\mu}(\mu, z) = \begin{cases} \Big[g'(\mu) - \frac{\alpha g(\mu)}{\mu^2}\big(Z_{\text{fin}} - z\big)\Big]\exp\Big(\frac{\alpha(Z_{\text{fin}}-z)}{\mu}\Big) & \text{if } (\mu, z) \in Q_-, \\ 0 & \text{if } (\mu, z) \in \{0\}\times[Z_{\text{ini}}, Z_{\text{fin}}], \\ \Big[f'(\mu) + \frac{\alpha f(\mu)}{\mu^2}\big(z - Z_{\text{ini}}\big)\Big]\exp\Big(-\frac{\alpha(z-Z_{\text{ini}})}{\mu}\Big) & \text{if } (\mu, z) \in Q_+. \end{cases}$$

A detailed proof can be done by observing that

$$\Big(\frac{\partial\psi}{\partial\mu}\Big)_{|_{Q_+}}(\mu, Z_{\text{ini}}) = f'(\mu), \quad \Big(\frac{\partial\psi}{\partial\mu}\Big)_{|_{Q_-}}(\mu, Z_{\text{fin}}) = g'(\mu),$$

and by using the inequalities

$$\Big|\frac{\partial\psi}{\partial\mu}(\mu, z)\Big| \leq \begin{cases} |g'(\mu)| + \alpha\big(Z_{\text{fin}} - Z_{\text{ini}}\big)\Big|\frac{g(\mu)}{\mu^2}\Big| & \text{if } \mu < 0, \\ 0 & \text{if } \mu = 0, \\ |f'(\mu)| + \alpha\big(Z_{\text{fin}} - Z_{\text{ini}}\big)\Big|\frac{f(\mu)}{\mu^2}\Big| & \text{if } \mu > 0, \end{cases}$$

valid for every $(\mu, z) \in Q$.

Summarizing, $\psi \in \mathrm{C}^1(Q)$ if, and only if, the following conditions are fulfilled:

1. $f \in \mathrm{C}^1\big([0,1]\big)$, $g \in \mathrm{C}^1\big([-1,0]\big)$, and
2. $\lim_{\mu\downarrow 0}\frac{f(\mu)}{\mu^2} = \lim_{\mu\uparrow 0}\frac{g(\mu)}{\mu^2} = 0$.

Equivalently, $\psi \in \mathrm{C}^1(Q)$ if, and only if, the following conditions are fulfilled:

1. $f \in \mathrm{C}^1\big([0,1]\big)$, $g \in \mathrm{C}^1\big([-1,0]\big)$, and
2. f and g are both twice differentiable at $\mu = 0$ (with the appropriate laterality), $f(0) = g(0) = 0$, $f'(0) = g'(0) = 0$ and $f''(0) = g''(0) = 0$.

Let us see the equivalence for f (it can be done for g analogously). Consider $f \in \mathrm{C}^1\big([0,1]\big)$.

Notice that $\lim_{\mu\downarrow 0}\frac{f(\mu)}{\mu^2}=0$ implies firstly $f(0)=0$ and secondly $f'(0)=\lim_{\mu\downarrow 0}\frac{f(\mu)-f(0)}{\mu}=\lim_{\mu\downarrow 0}\frac{f(\mu)}{\mu}=0$. Also, by L'Hôpital's rule,

$$0=2\lim_{\mu\downarrow 0}\frac{f(\mu)}{\mu^2}=\lim_{\mu\downarrow 0}\frac{f'(\mu)}{\mu}=\lim_{\mu\downarrow 0}\frac{f'(\mu)-f'(0)}{\mu}=f''(0).$$

Reciprocally, if $f(0)=f'(0)=f''(0)=0$, then, again by L'Hôpital's rule,

$$2\lim_{\mu\downarrow 0}\frac{f(\mu)}{\mu^2}=\lim_{\mu\downarrow 0}\frac{f'(\mu)}{\mu}=\lim_{\mu\downarrow 0}\frac{f'(\mu)-f'(0)}{\mu}=f''(0)=0.$$

Remarkably, the condition $f(0)=0$ is needed so that L'Hôpital's can be applied.

Class 2 analysis.

It can be proved that $\psi\in \mathrm{C}^2(Q)$ if, and only if, the following conditions are fulfilled:

1. $f\in \mathrm{C}^2\big([0,1]\big)$, $g\in \mathrm{C}^2\big([-1,0]\big)$, and
2. $\lim_{\mu\downarrow 0}\frac{f(\mu)}{\mu^4}=\lim_{\mu\uparrow 0}\frac{g(\mu)}{\mu^4}=0$.

Under these hypotheses, one has

$$\frac{\partial^2\psi}{\partial z^2}(\mu,z)=\begin{cases}\frac{\alpha^2 g(\mu)}{\mu^2}\exp\Big(\frac{\alpha(Z_{\mathrm{fin}}-z)}{\mu}\Big) & \text{if } (\mu,z)\in Q_-,\\ 0 & \text{if } (\mu,z)\in\{0\}\times[Z_{\mathrm{ini}},Z_{\mathrm{fin}}],\\ \frac{\alpha^2 f(\mu)}{\mu^2}\exp\Big(-\frac{\alpha(z-Z_{\mathrm{ini}})}{\mu}\Big) & \text{if } (\mu,z)\in Q_+,\end{cases}$$

$$\frac{\partial^2\psi}{\partial z\partial\mu}(\mu,z)=\frac{\partial^2\psi}{\partial\mu\partial z}(\mu,z)=$$

$$\begin{cases}\Big(\frac{\alpha g(\mu)}{\mu^2}-\frac{\alpha g'(\mu)}{\mu}+\frac{\alpha^2 g(\mu)}{\mu^3}\big(Z_{\mathrm{fin}}-z\big)\Big)\exp\Big(\frac{\alpha(Z_{\mathrm{fin}}-z)}{\mu}\Big) & \text{if } (\mu,z)\in Q_-,\\ 0 & \text{if } (\mu,z)\in\{0\}\times[Z_{\mathrm{ini}},Z_{\mathrm{fin}}],\\ \Big(\frac{\alpha f(\mu)}{\mu^2}-\frac{\alpha f'(\mu)}{\mu}-\frac{\alpha^2 f(\mu)}{\mu^3}\big(z-Z_{\mathrm{ini}}\big)\Big)\exp\Big(-\frac{\alpha(z-Z_{\mathrm{ini}})}{\mu}\Big) & \text{if } (\mu,z)\in Q_+,\end{cases}$$

and

$$\frac{\partial^2\psi}{\partial\mu^2}(\mu,z)=$$

$$\begin{cases}\Big[g''(\mu)+2\alpha\big(Z_{\mathrm{fin}}-z\big)\big(\frac{g(\mu)}{\mu^3}-\frac{g'(\mu)}{\mu^2}\big)+\alpha^2\big(Z_{\mathrm{fin}}-z\big)^2\frac{g(\mu)}{\mu^4}\Big]\times & \\ \qquad\exp\Big(\frac{\alpha(Z_{\mathrm{fin}}-z)}{\mu}\Big) & \text{if } (\mu,z)\in Q_-,\\ 0 & \text{if } (\mu,z)\in\{0\}\times[Z_{\mathrm{ini}},Z_{\mathrm{fin}}],\\ \Big[f''(\mu)+2\alpha\big(z-Z_{\mathrm{ini}}\big)\big(-\frac{f(\mu)}{\mu^3}+\frac{f'(\mu)}{\mu^2}\big)+\alpha^2\big(z-Z_{\mathrm{ini}}\big)^2\frac{f(\mu)}{\mu^4}\Big]\times & \\ \qquad\exp\Big(-\frac{\alpha(z-Z_{\mathrm{ini}})}{\mu}\Big) & \text{if } (\mu,z)\in Q_+,\end{cases}$$

with all the partial derivatives of second order continuous.

It is also true that $\psi\in \mathrm{C}^2(Q)$ if the following conditions are fulfilled:

1. $f \in \mathrm{C}^2\big([0,1]\big),\ g \in \mathrm{C}^2\big([-1,0]\big)$, and

2. f and g are both four times differentiable at $\mu = 0$ (with the appropriate laterality), $f(0) = g(0) = 0$, $f'(0) = g'(0) = 0$, $f''(0) = g''(0) = 0$, $f'''(0) = g'''(0) = 0$ and $f^{4)}(0) = g^{4)}(0) = 0$.

However, in contrast with the class 1 analysis, these two conditions are equivalent to the former ones only in case that both f and g be three times differentiable in a neighbourhood of $\mu = 0$. This is because $f \in \mathrm{C}^2\big([0,1]\big)$ and $\lim_{\mu\downarrow 0} \frac{f(\mu)}{\mu^4} = 0$ only imply that f is three times differentiable at $\mu = 0$ with $f(0) = f'(0) = f''(0) = f'''(0) = 0$, but not that f is three times differentiable in a neighbourhood of $\mu = 0$, which is necessary for f to be four times differentiable at $\mu = 0$.

Class k analysis.

It can be proved that $\psi \in \mathrm{C}^k(Q)$, with $k \in \mathbb{N}$, if, and only if, the following conditions are fulfilled:

1. $f \in \mathrm{C}^k\big([0,1]\big),\ g \in \mathrm{C}^k\big([-1,0]\big)$, and

2. $\lim_{\mu\downarrow 0} \frac{f(\mu)}{\mu^{2k}} = \lim_{\mu\uparrow 0} \frac{g(\mu)}{\mu^{2k}} = 0$.

The conditions above can be paraphrased as follows: *ψ is of class k if, and only if, f and g are of class k and flatter than μ^{2k} at $\mu = 0$.*

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

Direct method

Contents

3.1. The numerical scheme 24

3.1.1. Derivation of the numerical scheme 26

3.1.2. Description of the numerical scheme 27

3.2. Numerical results 29

3.2.1. Problems with known regular solutions 29

3.2.2. Examples from Kim and Tranquilli [21]. 32

3.3. Order and order* 38

One of the main results of the thesis is given in this Chapter, which is devoted to the direct method consisted of the even and the odd schemes and numerical results which will evince the order of the method. The odd scheme will be exploited in the Chapters 5 and 7.

3.1. The numerical scheme

From the viewpoint of the numerical scheme, variable z is to be interpreted as time, and variable μ as space. This mental abstraction is helpful for establishing the analogies with the classical finite difference schemes employed for solving the evolutive 1D heat equation. Accordingly, the expressions *initial* and *final condition* will be used to make reference to Equations *(2)* and *(3)*, respectively.

As in Chapter 2, let us call $Q_- = [-1, 0) \times [Z_{\text{ini}}, Z_{\text{fin}}]$, $Q_+ = (0, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}]$ and $Q_0 = \{0\} \times [Z_{\text{ini}}, Z_{\text{fin}}]$.

One could think from previous Chapter that we have a "final value problem" (FVP) determined by $g(\mu)$ on Q_- and an "initial vale problem" (IVP) determined by $f(\mu)$ on Q_+. However, such an assertion may be misleading, as these two problems are not independent except in the trivial situation $\sigma \equiv 0$, which is not our case.

Indeed, the presence of the diffusive term $\sigma \frac{\partial}{\partial \mu}\left[D(\mu)\frac{\partial \psi}{\partial \mu}\right]$ makes $\psi_{|Q_-}$ be dependent on $f(\mu)$ as well, and clearly a reciprocal comment holds for $\psi_{|Q_+}$. Naturally, this exchange of information between Q_- and Q_+ affects the values of the solution at the points of Q_0 in such a way that it is not possible to know them in advance.

We emphasize that, in case one desired to solve this problem (either on Q_- or on Q_+) in a "step by step" way, as it is done for solving the evolutive heat equation, the values on Q_0, which are unknown, would be required (see Figure 3.1).

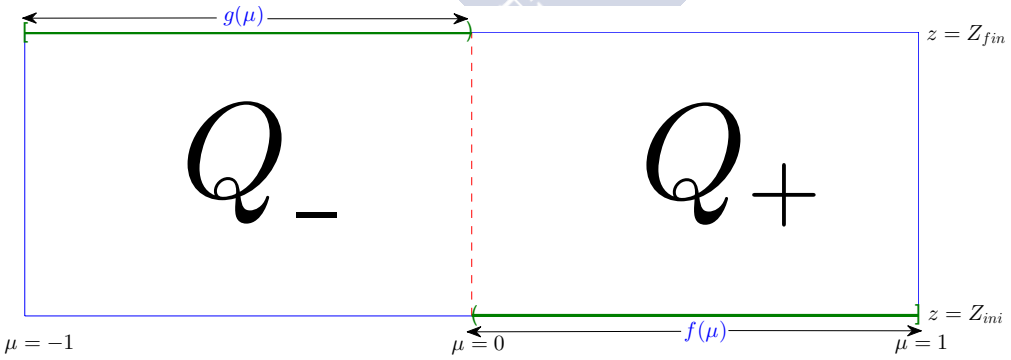


Figure 3.1: The green lines on the boundary of Q are marking out the set of points where the solution ψ is given by functions f and g.

On the other hand, it is also impossible to employ a "step by step" marching method on the whole Q, since $\psi_{|\{(\mu, Z_{\text{ini}}):\, \mu \in [-1,0]\}}$ is left to go forward and $\psi_{|\{(\mu, Z_{\text{fin}}):\, \mu \in [0,1]\}}$ is left to go backward.

Consequently, one must rather think

- either of using an iterative process starting with initial guess of the solution on Q_0,
- or of using a global scheme like that used for solving the two-dimensional Poisson equation, where the approximations at all mesh points (μ_i, z_n) are simultaneously obtained as the solution of a single large (but sparse) linear system.

We concentrate on the second approach. First case will be investigated in Chapter 5. Since it is also true that this is a problem with initial-final value structure, implicit numerical schemes should be able to solve it more robustly than explicit ones. Considering now that what it is written as implicit on the FVP half Q_- becomes explicit on the IVP half Q_+,[(d)] and vice versa, one concludes that a Crank-Nicolson-like scheme, possessing an implicit and an explicit part with the same weights, is a perfect candidate for being used on the whole grid without modifying the schemes's appearance. The details are as follows:

For given natural numbers I and N strictly greater than 1, let us consider the uniform meshes

$$\mu_i = -1 + (i-1)h \text{ for } i \in \{1, \ldots, I\}, \text{ with } h = \Delta\mu = \frac{2}{I-1} \tag{3.1}$$

and

$$z_n = Z_{\text{ini}} + (n-1)k \text{ for } k \in \{1, \ldots, N\}, \text{ with } k = \Delta z = \frac{Z_{\text{fin}} - Z_{\text{ini}}}{N-1}. \tag{3.2}$$

Hence,

$$\mu_1 = -1 < \mu_2 < \cdots < \mu_{I-1} < \mu_I = 1 \tag{3.3}$$

and

$$z_1 = Z_{\text{ini}} < z_2 < \cdots < z_{N-1} < z_N = Z_{\text{fin}}. \tag{3.4}$$

In the sequel, the following notations will be employed:

- $D(\mu) = 1 - \mu^2$.
- ψ_i^n will be approximation of the unknown solution ψ at the mesh point (μ_i, z_n) : $\psi_i^n \approx \psi(\mu_i, z_n) =: \overline{\psi}_i^n$.
- $z_{n+\frac{1}{2}} = z_n + \frac{k}{2}$.
- For a given function $\mathscr{A}$ of μ_i, $\overline{\mathscr{A}}_i = \mathscr{A}(\mu_i)$ and $\overline{\mathscr{A}}_{i\pm\frac{1}{2}} = \mathscr{A}(\mu_i \pm \frac{h}{2})$.
- For a given function $\mathscr{B}$ of (μ, z), $\overline{\mathscr{B}}_i^n = \mathscr{B}(\mu_i, z_n)$.

We shall below derive a numerical scheme of Crank-Nicolson type, hence of order $O(h^2) + O(k^2)$ for regular solutions. As to similar studies, we can mention that Vanaja, in reference [32], uses a finite difference scheme, but of order $O(h) + O(k)$ only, for solving the particular case in which $\alpha \equiv 0$, σ is a positive constant, and $W \equiv 0$. This authoress follows closely

[(d)] Another way to say the same thing: on Q_- an implicit scheme must be written in "forward form", while on Q_+ the same scheme must be written in "backward form".

the lines marked by herself and Kellogg in [33] for a slightly different kind of problems; in particular, an iterative fixed point procedure, starting with a guess of the solution on $\mu = 0$, is used to obtain the solution. Vanaja's z-discretization is of implicit Euler type and, as was mentioned above, it has to be written as "backward Euler" in order to be used on the forward parabolic part, i.e., on Q_+, and as "forward Euler" in order to be used on the backward parabolic part, i.e., on Q_-. Neither graphics nor numerical values of the solution are shown within the numerical results of [32], and hence comparison with these results is not possible.

Kim and Tranquilli, in reference [21], follow Morel [27] in that they choose as μ-nodes the I Gauss quadrature points in $(-1, 1)$, being I always even. These authors solve the particular case in which α and σ depend on z, but not on μ, and $W \equiv 0$. Kim and Tranquilli use an iterative method of fixed point type different from the one employed in [32]. The pictures within the numerical results of [21] can be used for graphical comparison, but it is not clear which is the order of the scheme with respect to h and k simultaneously.

Remarkably, $\mu = 0$ is always a Vanaja's node but never Kim-Tranquilli's node.

3.1.1. Derivation of the numerical scheme

Firstly one follows the idea of the Crank-Nicolson method for the heat equation to get

$$\mu_i \frac{\overline{\psi}_i^{n+1} - \overline{\psi}_i^n}{k} \overset{O(k^2)}{\approx} \mu_i \frac{\partial \psi}{\partial z}(\mu_i, z_{n+\frac{1}{2}}) \overset{O(k^2)}{\approx} \frac{\mu_i}{2}\Big(\frac{\partial \psi}{\partial z}(\mu_i, z_n) + \frac{\partial \psi}{\partial z}(\mu_i, z_{n+1})\Big). \tag{3.5}$$

Now, for $m \in \{n, n+1\}$,

$$\mu_i \frac{\partial \psi}{\partial z}(\mu_i, z_n) = \overline{W}_i^m - \overline{\alpha}_i^m \overline{\psi}_i^m + \overline{\sigma}_i^m \frac{\partial}{\partial \mu}\Big[D(\mu)\frac{\partial \psi}{\partial \mu}\Big]_{(\mu,z)=(\mu_i,z_m)} \tag{3.6}$$

according to Equation *(1)*, and finally the diffusive term is discretized as follows:

1. If $i = 1$, it is useful to notice that $\overline{D}_1 = 0$, which allows one to proceed as follows (see for instance [17, p. 207]:

$$\frac{\partial}{\partial \mu}\Big[D(\mu)\frac{\partial \psi}{\partial \mu}\Big]_{(\mu,z)=(\mu_1,z_m)} \overset{O(h^2)}{\approx} \frac{4\overline{D}_2 \frac{\partial \psi}{\partial \mu}(\mu_2, z_m) - \overline{D}_3 \frac{\partial \psi}{\partial \mu}(\mu_3, z_m)}{2h} \tag{3.7}$$

   and, for $r \in \{2, 3\}$, the standard centered scheme of two points gives

$$\frac{\partial \psi}{\partial \mu}(\mu_r, z_m) \overset{O(h^2)}{\approx} \frac{\overline{\psi}_{r+1}^m - \overline{\psi}_{r-1}^m}{2h}. \tag{3.8}$$

2. If $i \in \{2, \ldots, I-1\}$, the standard second-order centered formula gives

$$\frac{\partial}{\partial \mu}\Big[D(\mu)\frac{\partial \psi}{\partial \mu}\Big]_{(\mu,z)=(\mu_i,z_m)} \overset{O(h^2)}{\approx} \frac{\overline{D}_{i-\frac{1}{2}}\overline{\psi}_{i-1}^m - \Big(\overline{D}_{i-\frac{1}{2}} + \overline{D}_{i+\frac{1}{2}}\Big)\overline{\psi}_i^m + \overline{D}_{i+\frac{1}{2}}\overline{\psi}_{i+1}^m}{h^2}. \tag{3.9}$$

3. If $i = I$, the situation is analogous to that of $i = 1$. In this case, one takes advantage of the equality $\overline{D}_I = 0$ to write

$$\frac{\partial}{\partial\mu}\left[D(\mu)\frac{\partial\psi}{\partial\mu}\right]_{(\mu,z)=(\mu_I,z_m)} \overset{O(h^2)}{\approx} \frac{\overline{D}_{I-2}\frac{\partial\psi}{\partial\mu}(\mu_{I-2},z_m) - 4\overline{D}_{I-1}\frac{\partial\psi}{\partial\mu}(\mu_{I-1},z_m)}{2h} \tag{3.10}$$

and, for $r \in \{I-2, I-1\}$, *(3.8)* is used again.

3.1.2. Description of the numerical scheme

The above differentiation formulas suggest, after some reordering, the following scheme, which is well-defined if $I \geq 4$, I even, and $N \geq 2$. As it will be seen, a correction must be done in order to use the scheme for $I > 4$, I odd.

- For $(i,n) \in \{1\} \times \{1,\dots,N-1\}$,

$$\begin{aligned}\left(-\frac{\mu_1}{k}+\frac{\overline{\alpha}_1^n}{2}+\frac{\overline{\sigma}_1^n\overline{D}_2}{2h^2}\right)\psi_1^n+\left(-\frac{\overline{\sigma}_1^n\overline{D}_3}{8h^2}\right)\psi_2^n+\left(-\frac{\overline{\sigma}_1^n\overline{D}_2}{2h^2}\right)\psi_3^n+\\ \left(\frac{\overline{\sigma}_1^n\overline{D}_3}{8h^2}\right)\psi_4^n+\left(\frac{\mu_1}{k}+\frac{\overline{\alpha}_1^{n+1}}{2}+\frac{\overline{\sigma}_1^{n+1}\overline{D}_2}{2h^2}\right)\psi_1^{n+1}+\left(-\frac{\overline{\sigma}_1^{n+1}\overline{D}_3}{8h^2}\right)\psi_2^{n+1}+\\ \left(-\frac{\overline{\sigma}_1^{n+1}\overline{D}_2}{2h^2}\right)\psi_3^{n+1}+\left(\frac{\overline{\sigma}_1^{n+1}\overline{D}_3}{8h^2}\right)\psi_4^{n+1}=\frac{\overline{W}_1^n+\overline{W}_1^{n+1}}{2}.\end{aligned} \tag{3.11}$$

- For $(i,n) \in \{2,\dots,I-1\} \times \{1,\dots,N-1\}$,

$$\begin{aligned}\left(-\frac{\overline{\sigma}_i^n\overline{D}_{i-\frac{1}{2}}}{2h^2}\right)\psi_{i-1}^n+\left(-\frac{\mu_i}{k}+\frac{\overline{\alpha}_i^n}{2}+\frac{\overline{\sigma}_i^n\left(\overline{D}_{i-\frac{1}{2}}+\overline{D}_{i+\frac{1}{2}}\right)}{2h^2}\right)\psi_i^n+\\ \left(-\frac{\overline{\sigma}_i^n\overline{D}_{i+\frac{1}{2}}}{2h^2}\right)\psi_{i+1}^n+\left(-\frac{\overline{\sigma}_i^{n+1}\overline{D}_{i-\frac{1}{2}}}{2h^2}\right)\psi_{i-1}^{n+1}+\\ \left(\frac{\mu_i}{k}+\frac{\overline{\alpha}_i^{n+1}}{2}+\frac{\overline{\sigma}_i^{n+1}\left(\overline{D}_{i-\frac{1}{2}}+\overline{D}_{i+\frac{1}{2}}\right)}{2h^2}\right)\psi_i^{n+1}+\\ \left(-\frac{\overline{\sigma}_i^{n+1}\overline{D}_{i+\frac{1}{2}}}{2h^2}\right)\psi_{i+1}^{n+1}=\frac{\overline{W}_i^n+\overline{W}_i^{n+1}}{2}.\end{aligned} \tag{3.12}$$

- For $(i,n) \in \{I\} \times \{1,\dots,N-1\}$,

$$\begin{aligned}\left(\frac{\overline{\sigma}_I^n\overline{D}_{I-2}}{8h^2}\right)\psi_{I-3}^n+\left(-\frac{\overline{\sigma}_I^n\overline{D}_{I-1}}{2h^2}\right)\psi_{I-2}^n+\left(-\frac{\overline{\sigma}_I^n\overline{D}_{I-2}}{8h^2}\right)\psi_{I-1}^n+\\ \left(-\frac{\mu_I}{k}+\frac{\overline{\alpha}_I^n}{2}+\frac{\overline{\sigma}_I^n\overline{D}_{I-1}}{2h^2}\right)\psi_I^n+\left(\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-2}}{8h^2}\right)\psi_{I-3}^{n+1}+\\ \left(-\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-1}}{2h^2}\right)\psi_{I-2}^{n+1}+\left(-\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-2}}{8h^2}\right)\psi_{I-1}^{n+1}+\\ \left(\frac{\mu_I}{k}+\frac{\overline{\alpha}_I^{n+1}}{2}+\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-1}}{2h^2}\right)\psi_I^{n+1}=\frac{\overline{W}_I^n+\overline{W}_I^{n+1}}{2}.\end{aligned} \tag{3.13}$$

Correction for I odd. It is interesting to notice that counting the number of equations and the number of unknowns, which must coincide, gives a preponderant[(e)] position to even values I. The number of unknowns is $I \times N$, and the number of equations in the scheme above is $I \times (N-1)$; moreover, the initial and final conditions from Equations *(2)* and *(3)* provides one with I values when I is even, namely

$$\psi_i^1 = \overline{f}_i, \text{ for } i = \frac{I}{2}+1, \ldots, I \text{ and } \psi_i^N = \overline{g}_i \text{ for } i = 1, \ldots, \frac{I}{2}, \tag{3.14}$$

and with only $I-1$ values when I is odd, namely

$$\psi_i^1 = \overline{f}_i \text{ for } i = \frac{I+3}{2}, \ldots, I \text{ and } \psi_i^N = \overline{g}_i \text{ for } i = 1, \ldots, \frac{I-1}{2}. \tag{3.15}$$

As a result, the number of equations equals the number of unknowns when I is even, but there is one equation left when I is odd.

In conclusion:

- If I is even, the linear system is closed.
- If I is odd, an extra equation is needed to close the system. Observe that $\mu_{\frac{I+1}{2}} = 0$. As a first approach (which will be improved below) one could think of using either the *numerical initial condition*

$$\psi^1_{\frac{I+1}{2}} = f(0) \tag{3.16}$$

or the *numerical final condition*

$$\psi^N_{\frac{I+1}{2}} = g(0). \tag{3.17}$$

We remark that only one of these two equations must be chosen. This approach, even when it has been tested with acceptable results, is asymmetric, since in fact the imposition of both conditions *(3.16)* and *(3.17)*, and not only one of them, would be desirable. That is the reason why it is improved below. It is natural to proceed as follows. The idea is to discard the $N-1$ equations corresponding to the choice $i = \frac{I+1}{2}$ in Equation *(3.12)*,[(f)] to impose conditions *(3.16)* and *(3.17)*, and to consider a new set of $N-2$ equations which will close the linear system. In order to obtain this new set of equations, simply notice that Equation *(1)* becomes, for $\mu = 0$,

$$\alpha(0,z)\psi(0,z) - \sigma(0,z)\frac{\partial^2 \psi}{\partial \mu^2}(0,z) = W(0,z) \text{ for } z \in [Z_{\text{ini}}, Z_{\text{fin}}]. \tag{3.18}$$

Now Equation *(3.18)* suggests, for $i^* = \frac{I+1}{2}$ and $n = \{2, \ldots, N-1\}$,

$$\overline{\alpha}^n_{i^*}\psi^n_{i^*} - \overline{\sigma}^n_{i^*}\frac{\psi^n_{i^*-1} - 2\psi^n_{i^*} + \psi^n_{i^*+1}}{h^2} = \overline{W}^n_{i^*} \tag{3.19}$$

or, equivalently,

$$\left(-\frac{\overline{\sigma}^n_{i^*}}{h^2}\right)\psi^n_{i^*-1} + \left(\overline{\alpha}^n_{i^*} + \frac{2\overline{\sigma}^n_{i^*}}{h^2}\right)\psi^n_{i^*} + \left(-\frac{\overline{\sigma}^n_{i^*}}{h^2}\right)\psi^n_{i^*+1} = \overline{W}^n_{i^*}. \tag{3.20}$$

[(e)] The adjective "preponderant" is used in relation to the well-posedness of the linear system and it has nothing to do with accuracy. In fact, it will be demonstrated later that it is advantageous to take odd values of I.

[(f)] There is no reason to prefer removing the first one versus the last one of these. Therefore, we eliminate all of them in order to pursue other reasoning.

This new set of equations is also based upon second order approximations of the derivatives, and hence it is not supposed to spoil the order $O(h^2)+O(k^2)$ of the method. This observation will be supported by numerical evidence.

So, to summarize, when I is odd the $N-1$ equations corresponding to the choice $i = \frac{I+1}{2}$ in Equations *(3.12)* are discarded, and the $N-2$ new equations given by *(3.20)* plus Equations *(3.16)* and *(3.17)* are considerd instead.

Hereinafter, we shall use the expression "even scheme" to mean the use of I even, and "odd scheme" to mean the use of I odd.

3.2. Numerical results

Both the even and the odd scheme have been implemented in MATLAB® (R2012b). As it was explained above, the approximate values ψ_i^n are obtained by solving a linear sparse system of dimension $I \times N$. In this Section:

- $E_{\text{abs}}(Q) = \max_Q|\psi_{\text{grid}} - \psi|$, where ψ_{grid} is representing the approximate solution and the maximum is taken over the set of all nodes.
- $D_{\text{abs}}(Q) = \max_Q|\psi_{\text{coarse}} - \psi_{\text{fine}}|$ is the maximum punctual difference between a coarse-fine embedded pair of numerical approximations: the $(2I-1, 2N-1)$ fine grid is built by roughly doubling the (I, N) coarse grid in each variable. The maximum is taken over all common nodes, that is, over all the nodes of the coarse grid.
- The (μ, z) column within the tables reports the grid point where the value of $E_{\text{abs}}(Q)$ or $D_{\text{abs}}(Q)$ on the left is attained.

Numerical results show that the (odd or even) scheme converges with the expected order 2 in both μ and z. Notice that this is an experimental assertion, as no numerical analysis has been carried out. The order and order* within the tables have been truncated after performing their computation with more decimals than those present in the error columns (consequently, if one does the operations employing the numbers as they occur in the tables, the result will differ slightly). In all our experiments, the odd scheme has performed better than (or as good as) the even scheme in the sense of computational time.

3.2.1. Problems with known regular solutions

Here, the adjective "regular" is being used as synonymous of "belonging to $\mathrm{C}^\infty(Q)$". Test cases with known regular solution are useful to check scheme's convergence and order. They are easily derived thanks to the presence of the source term W; the idea is to fix, freely, Z_{ini}, Z_{fin}, α and σ, as well as a function $\psi \in \mathrm{C}^\infty(Q)$, which is going to be the exact solution. The data functions left, namely, W, f and g are computed from Equations *(1)–(3)*.

The following facts, which by the way are expected from the discretizations employed,

will be observed for regular ψ (assertions about "the scheme" are valid for both the even and the odd scheme):

(1) The scheme solves the problem exactly if $\psi(\cdot, z)$ is a polynomial of degree ≤ 1 in μ and $\psi(\mu, \cdot)$ is a polynomial of degree ≤ 2 in z. See Table 3.2.1.

(2) The scheme is exact with respect to z and converges with order $O(h^2)$ for constant k if $\psi(\mu, \cdot)$ is a polynomial of degree ≤ 2 in z. See Table 3.2.2.

(3) The scheme is exact with respect to μ and converges with order $O(k^2)$ for constant h if $\psi(\cdot, z)$ is a polynomial degree ≤ 1 in μ. See Table 3.2.3.

(4) The scheme converges with order $O(h^2) + O(k^2)$. See Tables 3.2.4 and 3.2.5.

In the examples below, we consider $Z_{\text{ini}} = 0$, $Z_{\text{fin}} = 1$ and different combinations of

$$\alpha \in \{0, 1, |\sin(12\mu z)|, 2 + \sin(12\mu z)\} \tag{3.21}$$

and

$$\sigma \in \{1, 1 + \sin(12\mu z)\cos(12\mu z)\}. \tag{3.22}$$

Test case with known regular #1.

If the exact solution is

$$\psi(\mu, z) = \mu z^2, \tag{3.23}$$

then the method is exact, and only round-off errors occur. For all possible combinations of *(3.21)* and *(3.22)* the maximum error is less than 10^{-15} when $I = 11$ and $N = 10$. Table 3.2.1 shows the error for one of these combinations. Notice that round-off error is added when I and N increase.

$(I,\ N)$	$E_{\text{abs}}(Q)$
$(11, 10)$	3.88×10^{-16}
$(101, 91)$	5.22×10^{-15}
$(1001, 901)$	2.50×10^{-13}

Table 3.2.1: Numerical results for the test case with exact solution *(3.23)* for $\alpha(\mu, z) = |\sin(12\mu z)|$ and $\sigma(\mu, z) = 1 + \sin(12\mu z)\cos(12\mu z)$.

Test case with known regular #2.

If the exact solution is

$$\psi(\mu, z) = \mu^2 z^2, \tag{3.24}$$

then the method is exact with respect to z, and hence high accuracy can be reached by refining only the μ-mesh, that is, by augmenting I while maintaining N constant. In this case, the effect of increasing the value of N is to add round-off errors, and hence accuracy is not improved. See Table 3.2.2, where the order $O(h^2)$ is apparent. As expected from the form of ψ, accuracy is not improved when N is increased in the last row of this table.

The same behavior is observed every time that $\psi(\mu,\cdot)$ is a polynomial of degree ≤ 2 in z: for instance, $\psi(\mu,z)=\mu^3(3+2\mu z-z^2)$ or $\psi(\mu,z)=(z-3z^2)\sin\mu$.

A complementary case holds every time that $\psi(\cdot,z)$ is a polynomial of degree ≤ 1 in μ.

If the exact solution is, let us say,

$$\psi(\mu,z)=\mu z^3 \tag{3.25}$$

or

$$\psi(\mu,z)=(1+\mu\cos z)\sin z,$$

then the order $O(k^2)$ is achieved by increasing N while maintaining I constant. Results for a particular case can be seen in Table 3.2.3. Observe that, despite the remarkable refinement of the μ-mesh, the error in the last row does not diminish in a significant way; in other words, accuracy is not improved when I is increased.

(I,N)	$E_{\text{abs}}(Q)$	(μ,z)	order
$(11,10)$	1.26×10^{-2}	$(-1,\ 0.888...)$	
$(33,10)$	7.55×10^{-4}	$(-1,\ 0.888...)$	$\frac{2\ln(\frac{126}{7.55})}{\ln(10)}=2.447$
$(101,10)$	7.83×10^{-5}	$(0.8,1)$	1.967
$(321,10)$	8.29×10^{-6}	$(0.85652...,1)$	1.951
$(1001,10)$	8.72×10^{-7}	$(0.882,1)$	1.956
$(1001,901)$	9.23×10^{-7}	$(0.808,1)$	

Table 3.2.2: Numerical results for the test case with exact *(3.24)*, for $\alpha(\mu,z)=|\sin(12\mu z)|$ and $\sigma(\mu,z)=1+\sin(12\mu z)\cos(12\mu z)$.

(I,N)	$E_{\text{abs}}(Q)$	(μ,z)	order
$(11,10)$	2.81×10^{-3}	$(-1,0)$	
$(11,29)$	2.99×10^{-4}	$(-1,0)$	$\frac{2\ln(\frac{28.1}{2.99})}{\ln(10)}=1.949$
$(11,91)$	2.90×10^{-5}	$(-1,0)$	2.025
$(11,281)$	3.00×10^{-6}	$(-1,0)$	1.917
$(11,901)$	2.90×10^{-7}	$(-1,0)$	2.028
$(1001,901)$	2.87×10^{-7}	$(0.808,1)$	

Table 3.2.3: Numerical results for the test case with exact solution *(3.25)* $\psi(\mu,z)=\mu z^3$, for $\alpha(\mu,z)=|\sin(12\mu z)|$ and $\sigma(\mu,z)=1+\sin(12\mu z)\cos(12\mu z)$.

Test case with known regular #3.

If the exact solution is

$$\psi(\mu,z)=\ln(2+\mu^2+z^3), \tag{3.26}$$

then the method is not exact with respect to any of the variables, and hence I and N have to be simultaneously increased to improve accuracy with order 2. The comparison between the exact and numerical solutions as well as the order of convergence are shown in Table 3.2.4. Table 3.2.5, which compares the numerical solution on different meshes, shows that the value of order*, computed from $D_{\text{abs}}(Q)$, is a good indicator of the order.

(I, N)	$E_{\text{abs}}(Q)$	(μ, z)	order
$(11, 10)$	7.05×10^{-3}	$(-1, 0)$	
$(33, 29)$	5.78×10^{-4}	$(-0.0625, 0.17857)$	$\frac{2\ln(\frac{70.5}{5.78})}{\ln(10)} = 2.173$
$(101, 91)$	5.87×10^{-5}	$(-0.04, 0.16666...)$	1.986
$(321, 281)$	5.72×10^{-6}	$(-0.04375, 0.15714...)$	2.022
$(1001, 901)$	5.85×10^{-7}	$(-0.04, 0.15888...)$	1.981

Table 3.2.4: Numerical results for the test case with exact solution *(3.26)* for $\alpha(\mu, z) = |\sin(12\mu z)|$ and $\sigma(\mu, z) = 1 + \sin(12\mu z)\cos(12\mu z)$.

(I, N)	$(2I-1, 2N-1)$	$D_{\text{abs}}(Q)$	(μ, z)	order*
$(21, 21)$	$(41, 41)$	1.11×10^{-3}	$(-0.1,\ 0.2)$	
$(41, 41)$	$(81, 81)$	2.77×10^{-4}	$(-0.05, 0.175)$	$\frac{\ln(\frac{11.1}{2.77})}{\ln(2)} = 2.006$
$(81, 81)$	$(161, 161)$	6.87×10^{-5}	$(-0.025,\ 0.1625)$	2.010
$(161, 161)$	$(321, 321)$	1.71×10^{-5}	$(0.025, 0.16875)$	2.004
$(321, 321)$	$(641, 641)$	4.27×10^{-6}	$(-0.03125, 0.165625)$	2.002
$(641, 641)$	$(1281, 1281)$	1.07×10^{-6}	$(-0.03125,\ 0.0165625)$	2.0009
$(1281, 1281)$	$(2561, 2561)$	2.67×10^{-7}	$(-0.03125, 0.16484375)$	2.0005

Table 3.2.5: Numerical results for the test case with exact solution *(3.26)* for $\alpha(\mu, z) = |\sin(12\mu z)|$ and $\sigma(\mu, z) = 1 + \sin(12\mu z)\cos(12\mu z)$.

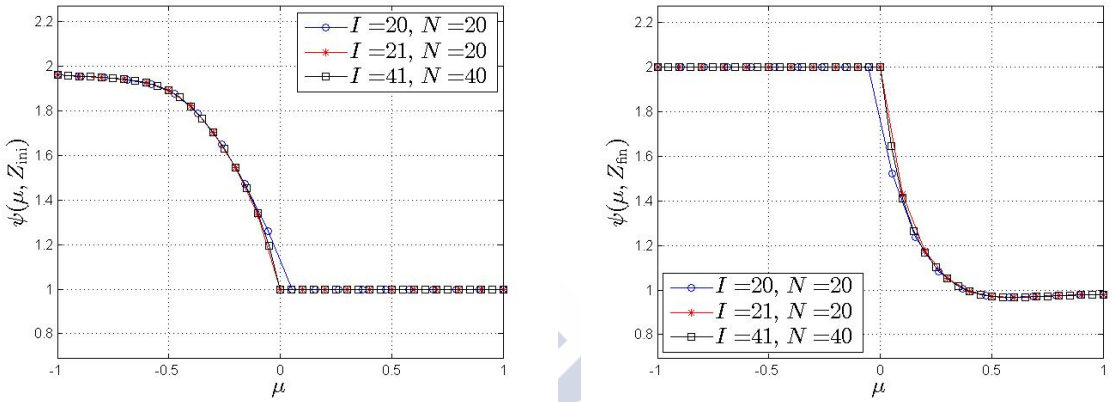
3.2.2. Examples from Kim and Tranquilli [21].

Kim and Tranquilli [21] employ the Fokker-Planck equation in the 1D slab motivated by the modeling of light propagation in biological tissue. They show some plots that can be used for comparison with our results. As to numerical results, we cannot compare it is because in the paper [21], numerical results are not presented.

Kim-Tranquilli's problem #1.

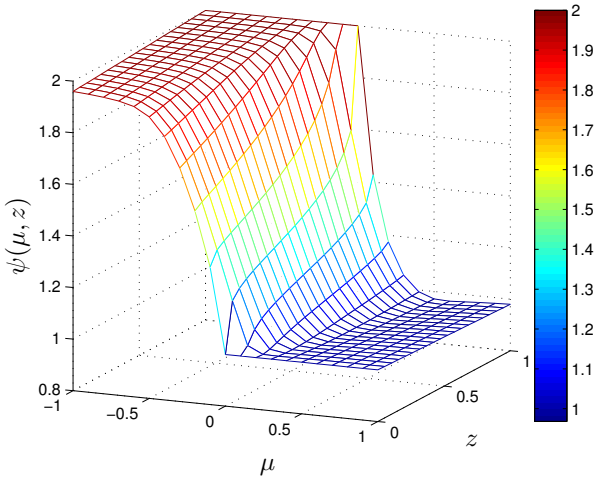
Problem (55a)-(55c) in reference [21] is considered: $Z_{\text{ini}} = 0$, $Z_{\text{fin}} = 1$, $\alpha(\mu, z) = 0.02$, $\sigma(\mu, z) = 0.01$, $f(\mu) = 1$, $g(\mu) = 2$, $W(\mu, z) = 0$.

Figure 3.2 shows that it is sufficient to take $I \in \{20, 21\}$, $N = 20$ in order to obtain a good agreement with the graphics in [21].



(a) At $z = Z_{\text{ini}} = 0$.

(b) At $z = Z_{\text{fin}} = 1$.



(c) Complete solution $I = 21, N = 20$.

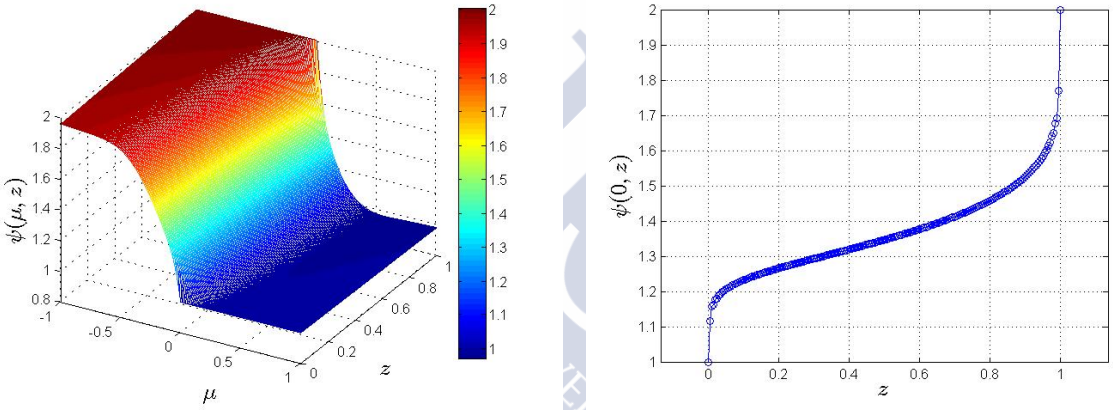
Figure 3.2: Approximate solution of Kim-Tranquilli's problem #1 obtained with the even and odd schemes for different meshes.

Numerical experiments performed with finer grids (see Figure 3.3) allow conjecturing that:

- ψ is continuous on Q.
- $\dfrac{\partial \psi}{\partial \mu}$ does not exist at $(0, Z_{\text{ini}})$ and at $(0, Z_{\text{fin}})$.
- $\lim_{z \downarrow Z_{\text{ini}}} \dfrac{\partial \psi}{\partial z}(0, z) = \lim_{z \uparrow Z_{\text{fin}}} \dfrac{\partial \psi}{\partial z}(0, z) = +\infty$.

(I, N)	$\widehat{E}_{\text{abs}}(Q)$	(μ, z)	order**
$(21, 21)$	6.48×10^{-2}	$(-0.1, 0.95)$	
$(41, 41)$	9.57×10^{-2}	$(0.05, 1)$	$\frac{\ln(\frac{648}{957})}{\ln(2)} = -0.56$
$(81, 81)$	9.62×10^{-2}	$(0.025, 1)$	-0.008
$(161, 161)$	7.65×10^{-2}	$(0.025, 1)$	0.33
$(321, 321)$	5.52×10^{-2}	$(0.025, 1)$	0.47
$(641, 641)$	4.65×10^{-2}	$(-0.0125, 0.9984375)$	0.25
$(1281, 1281)$	4.11×10^{-2}	$(-0.009375, 0.99921875)$	0.18

Table 3.2.7: Numerical results for Kim-Tranquilli's problem #1. $\widehat{E}_{\text{abs}}(Q)$ measures the error with respect to a reference solution computed for $I = N = 2561$.



(a) Complete solution. (b) At $\mu = 0$.

Figure 3.3: Approximate solution of Kim-Tranquilli's problem #1 obtained with the odd scheme for $I = 201$, $N = 200$.

(I, N)	$(2I-1, 2N-1)$	$D_{\text{abs}}(Q)$	(μ, z)	order*
$(21, 21)$	$(41, 41)$	3.34×10^{-2}	$(-0.1, 0.95)$	
$(41, 41)$	$(81, 81)$	6.17×10^{-2}	$(-0.05, 0.975)$	$\frac{\ln(\frac{334}{617})}{\ln(2)} = -0.89$
$(81, 81)$	$(161, 161)$	6.47×10^{-2}	$(-0.025, 0.98)$	-0.07
$(161, 161)$	$(321, 321)$	5.65×10^{-2}	$(-0.0125, 0.99375)$	0.19
$(321, 321)$	$(641, 641)$	5.16×10^{-2}	$(-0.0125, 0.996875)$	0.13
$(641, 641)$	$(1281, 1281)$	4.59×10^{-2}	$(-0.0125, 0.9984375)$	0.17

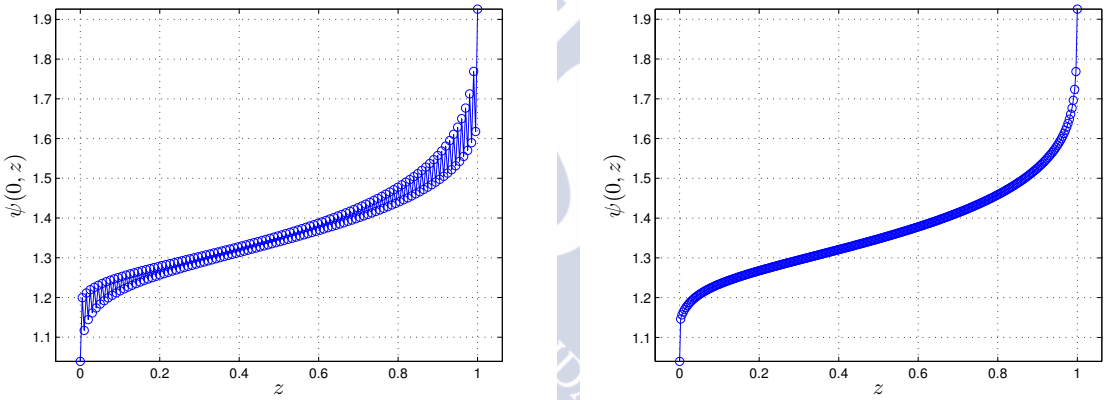
Table 3.2.6: Numerical results for Kim-Tranquilli's problem #1.

As expected from this analysis, convergence of order 2 is not observed and the numerical solution is less accurate, though no bad, near the singularities $\{(0, Z_{\text{ini}}), (0, Z_{\text{fin}})\}$. So far, as order 2 must be seen when the solution is regular, the lack of order 2 supports the idea, as do the plots in Figures 3.2 and 3.3, that the solution of this problem is not regular.

Numerical results shown in Table 3.2.6 are corroborated by those in Table 3.2.7. The quantity $\widehat{E}_{\text{abs}}(Q)$ in Table 3.2.7 is comparing the numerical solutions with a reference solution computed for $I = N = 2561$, once more by means of the maximum error over the set of common nodes.

Differences between the even and the odd scheme.

Plots in Figures 3.2(a) and 3.2(b) show similar results for the even and the odd schemes. However, there exist remarkable differences in favor of the odd one when the grid is refined. Indeed, results obtained with the even scheme easily suffer from spurious oscillations near $(0, Z_{\text{ini}})$ and $(0, Z_{\text{fin}})$, and in the worst cases these instabilities propagate along the vicinity of the whole segment $\mu = 0$,[(h)] while none of these problems arise when using the odd scheme. Figure 3.4(a), which must be contrasted with Figure 3.3(b), manifests this phenomenon. Figure 3.4(b) shows that instabilities tend to disappear when refining the z grid. Since $\mu = 0$ is not a node when I is even, the values at $\mu = 0$ have been computed as the arithmetic mean of the values at the nodes $\mu_{\frac{I}{2}} = -\frac{h}{2}$ and $\mu_{\frac{I}{2}+1} = \frac{h}{2}$.



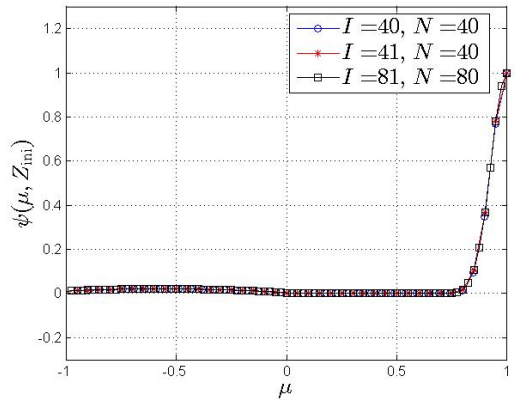
(a) $I = 200$, $N = 200$. (b) $I = 200$, $N = 1200$.

Figure 3.4: Approximate solution at $\mu = 0$ of Kim-Tranquilli's problem #1 obtained with the even scheme.

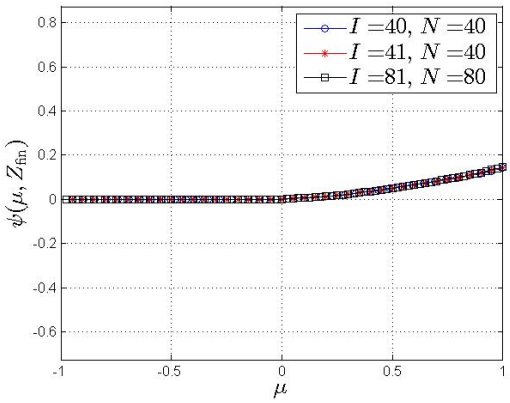
Kim-Tranquilli's problem #2.

Problem (34) with data (57)-(60) in reference [21] is considered: $Z_{\text{ini}} = 0$, $Z_{\text{fin}} = 1$, $\alpha(\mu, z) = 0.01{+}5\text{e}^{-500(z-0.6)^2}$, $\sigma(\mu, z) = \frac{1}{2}(0.01{+}5\text{e}^{-500(z-0.4)^2})$, $f(\mu) = \text{e}^{-100(\mu-1)^2}$, $g(\mu) = 0$, $W(\mu, z) = 0$. Once more, it is not necessary to use a very fine grid ($I \in \{40, 41\}$, $N = 40$ suffices) in order to obtain a good agreement of our Figure 3.5(c) with Figure 5 in reference [21].

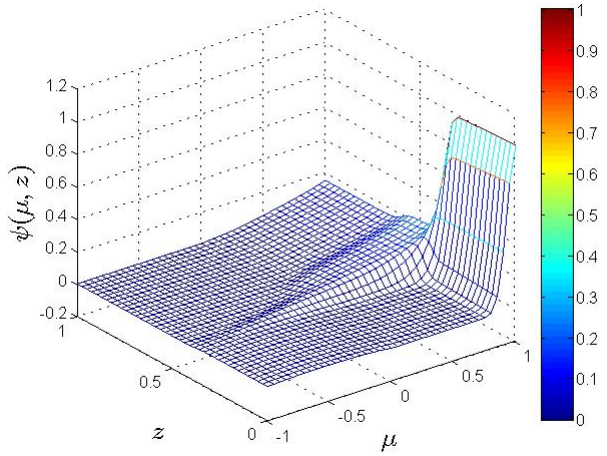
[(h)] These drawbacks of the even scheme for this problem could be corrected, in all of our trials, by taking $N \sim 6I$.



(a) At $z = Z_{\text{ini}} = 0$.



(b) At $z = Z_{\text{fin}} = 1$.



(c) Complete solution for $I = 41, N = 40$.

Figure 3.5: Approximate solution of Kim-Tranquilli's problem #2 obtained with the even and the odd schemes for different meshes.

As the authors say in [21], this problem models a plane wave normally incident, at $z = 0$, on a plane-parallel slab of tissue having unit thickness. Condition $f(\mu) = \mathrm{e}^{-100(\mu-1)^2}$ is regularizing the plane wave, mathematically modeled with a Dirac delta, by means of a narrow Gaussian, and condition $g(\mu) = 0$ means that no light enters the slab at $z = 1$. The slab of tissue has an absorbing inhomogeneity given by α and a scattering inhomogeneity given by 2σ.

Spurious oscillations do not occur when solving this problem with the even scheme and the results are comparable to those obtained with the odd scheme. Tables 3.2.8 and 3.2.9 report the numerical results. The quantity $\widehat{E}_{\text{abs}}(Q)$ in Table 3.2.9 has the same meaning as in Table 3.2.7. Notice that the order 2 disappears at the sixth and seventh rows, where round-off errors should not be spoiling yet the approximations. This fact seems to be indicating lack of regularity of ψ to an extent that cannot be specified without further theoretical research. We want to emphasize that by analyzing the order one can get information about the regularity, which, as it happens for this problem, might be imperceptible in the graphics.

(I, N)	$(2I-1, 2N-1)$	$D_{\text{abs}}(Q)$	(μ, z)	order*
$(21, 21)$	$(41, 41)$	3.67×10^{-2}	$(1, 0.35)$	
$(41, 41)$	$(81, 81)$	4.23×10^{-2}	$(1, 0.35)$	$\frac{\ln(\frac{367}{423})}{\ln(2)} = -0.20$
$(81, 81)$	$(161, 161)$	1.64×10^{-2}	$(1, 0.3375)$	1.37
$(161, 161)$	$(321, 321)$	3.53×10^{-3}	$(1, 0.3375)$	2.21
$(321, 321)$	$(641, 641)$	7.98×10^{-4}	$(0.0125, 0.003125)$	2.15
$(641, 641)$	$(1281, 1281)$	7.11×10^{-4}	$(0.009375, 0.0015625)$	0.16
$(1281, 1281)$	$(2561, 2561)$	6.33×10^{-4}	$(0.00625, 0.0078125)$	0.17

Table 3.2.8: Numerical results for Kim-Tranquilli's problem #2.

(I, N)	$\widehat{E}_{\text{abs}}(Q)$	$(\mu,\ z)$	order**
$(21, 21)$	9.35×10^{-2}	$(1, 0.35)$	
$(41, 41)$	5.67×10^{-2}	$(1, 0.35)$	$\frac{\ln(\frac{935}{567})}{\ln(2)} = 0.72$
$(81, 81)$	2.06×10^{-2}	$(1, 0.3375)$	1.46
$(161, 161)$	4.24×10^{-3}	$(1, 0.3375)$	2.28
$(321, 321)$	8.50×10^{-4}	$(-0.01875,\ 0)$	2.32
$(641, 641)$	7.17×10^{-4}	$(0.009375, 0.0015625)$	0.24
$(1281, 1281)$	6.33×10^{-4}	$(0.00625, 0.0078125)$	0.18

Table 3.2.9: Numerical results for Kim-Tranquilli's problem #2. $\widehat{E}_{\text{abs}}(Q)$ measures the error with respect to a reference solution computed for $I = N = 2561$.

Computing time.

Table 3.2.10 shows the sum of the time employed in defining the matrix and the second member, plus the time spent in solving the linear system. These times rely on a smart MATLAB® implementation.

Ultimately, what one does is to solve a sparse linear system of order $I \times N$, with a degree of sparsity that can be easily derived from the scheme. Hence, there is nothing essentially new in these computing times. We display them so that one can immediately judge the scheme's performance in a particular situation. To provide an example, if a refinement $I = 101$, $N = 100$ is good enough for the purposes at hand, then one can solve of the order of 630 Fokker-Planck problems like *(1)–(3)* in about 1 min by using a present-day PC.

I	N	Time (s)
$\{10, 11\}$	100	$\{0.022, 0.022\}$
$\{10, 11\}$	1000	$\{0.083, 0.096\}$
$\{100, 101\}$	100	$\{0.101, 0.095\}$
$\{100, 101\}$	1000	$\{0.559, 0.537\}$
$\{1000, 1001\}$	100	$\{0.614, 0.602\}$
$\{1000, 1001\}$	1000	$\{15.830, 13.600\}$
2561	2561	636.432

Table 3.2.10: Values of computing times by using a personal computer with an Intel® Core™ i7-4790 3.60GHz processor.

3.3. Order and order*

Wherever we say "the scheme" without further specification, it must be understood that both the even and the odd scheme are being considered. When we say "a scheme", it must be understood that we are talking about any numerical scheme for solving the problem *(1)–(3)*, and that scheme might be one of our schemes or not.

We have already defined $D_{\text{abs}}(Q)$ and $E_{\text{abs}}(Q)$ in Section 3.2 but, for the sake of language accuracy, we must redefine both concepts in a finer way. Both "abs" and "Q" will be removed for reasons of economy.

The error $E(h, k)$ defined below depends also on the exact solution ψ through some of its derivatives, but there is no need to make this dependence explicit for the purposes at hand. Notice, by the way, that the definition does not demand regularity of ψ.

Definition 3.3.1. *Let ψ be the exact solution, consider an (h, k) grid and let ψ_g be the numerical solution obtained for this grid. Then the real number $E(h, k)$ is defined as*

$$E(h, k) = \max_g |\psi_g - \psi|,$$

where the maximum is taken over all nodes.

In the forthcoming Definition 3.3.2, the adjective "smooth" is used with the meaning of "belonging to $C^\infty(Q)$". Even when this interpretation might be weakened, the key point in what regards this work is that the order of convergence is not observed when ψ is not regular enough.

The following definition of order has been thought for a numerical method for solving the problem *(1)–(3)* based on formulas that compute exactly

- $\frac{\partial}{\partial \mu}\left[(1-\mu^2)\frac{\partial \psi}{\partial \mu}\right]$ when, for every z, $\psi(\cdot, z)$ is a polynomial of degree $\leq p-1$ in μ, and
- $\frac{\partial \psi}{\partial z}$ when, for every μ, $\psi(\mu, \cdot)$ is a polynomial of degree $\leq q$ in z.

Notice that $\psi(\cdot, z)$ is a polynomial of degree $\leq p-1$ in μ if, and only if, $(1-\mu^2)\frac{\partial \psi}{\partial \mu}$ is a polynomial of degree $\leq p$ in μ.

Definition 3.3.2. *(*order*) Let p and q be two natural numbers. We say that a numerical scheme in the conditions above is of order $O(h^p)+O(k^q)$ if p and q are the largest real numbers such that*

$$E(h,k) = O(h^p) + O(k^q) \text{ when } \psi \text{ is smooth}$$

and, moreover, the following three assertions are simultaneously true:

1. p is the largest real number satisfying $E(h,k) = O(h^p)$ for any smooth ψ such that,

$$\text{for every } \mu, \psi(\mu, \cdot) \text{ is a polynomial of degree} \leq q \text{ in } z. \tag{3.27}$$

2. q is the largest real number satisfying $E(h,k) = O(k^q)$ for any smooth ψ such that,

$$\text{for every } z, \psi(\cdot, z) \text{ is a polynomial of degree} \leq p-1 \text{ in } \mu. \tag{3.28}$$

3. $E(h,k) = 0$ when ψ satisfies conditions *(3.27)* and *(3.28)*.

For simplicity, the case $p = q$ can be referred to as *order p*.

Remark 3.3.1. *In a more extended way, one can equally say that the scheme has order p with respect to h and order q with respect to k.*

Remark 3.3.2. *As usual when defining the order of convergence, the effect of rounding errors is neglected.*

Typically,[(i)] when a scheme has got order p, then, for any fixed number $c \in (0, \infty)$,

$$E(ch, ck) = c^p E(h,k) \text{ asymptotically}, \tag{3.29}$$

where "asymptotically" means that the equality is not a true equality, but it tends to become an equality when h and k tend to zero. So, $E(h,k)$ is somewhat a homogeneous function of degree p. Consequently, the value of p can be obtained from

$$p = \lim_{h \downarrow 0, k \downarrow 0} \frac{\ln\left(\frac{E(h,k)}{E(ch,ck)}\right)}{\ln c}, \tag{3.30}$$

being c any fixed positive number different from 1.

However, applying this reasoning for obtaining p experimentally may be treacherous. Let us explain further this point. Imagine that we have a scheme of order $O(h^p)+O(k^q)$ with, let us say,

$$E(h,k) = C_1 h^p + C_2 k^q. \tag{3.31}$$

Since in numerical experiments it is customary to take $k = mh$ for some positive constant m, the direct application of *(3.30)* gives an order equal to $\min\{p, q\}$, which is not true unless

[(i)] Equation *(3.29)* cannot be demonstrated from Definition 3.3.2.

$p = q$. In other words, formula *(3.30)* should not be employed until one knows by some other means that $p = q$.

We propose now a general procedure that can be used for any values of p and q. For a scheme of order $O(h^p) + O(k^q)$ one should write

$$E\left(c^{\frac{1}{p}}h, c^{\frac{1}{q}}k\right) = cE(h,k) \text{ asymptotically} \tag{3.32}$$

instead of Equation *(3.29)*, but, except when $p = q$ (which is not known *a priori*), this identity is of no use for obtaining p and q. However, the value of p can be obtained from $E(ch,k) = c^p E(h,k)$ for fixed k by taking a smooth function ψ satisfying *(3.27)*, and the value of q from $E(h,ck) = c^q E(h,k)$ for fixed h by taking a smooth function ψ satisfying *(3.28)*. Once p and q are known, the one must check *(3.29)* or *(3.32)* in order to conclude that the method is indeed of order p or of order $O(h^p) + O(k^q)$.

In fact, slightly stronger assertions than those used above usually hold: one can write $E\big(c(h,k)^{\frac{1}{p}}h), c(h,k)^{\frac{1}{q}}k\big)$ instead of $E\big(c^{\frac{1}{p}}h, c^{\frac{1}{q}}k\big)$, $E\big(c(h,k)h, c(h,k)k\big)$ instead of $E\big(ch, ck\big)$, and $\ln C$ instead of $\ln c$ provided that $c(h,k)$ is a function of h and k such that the limit

$$\lim_{h\downarrow 0, k\downarrow 0} c(h,k) = C \tag{3.33}$$

exists in $(0,\infty) \setminus \{1\}$.

Theorem 3.3.1. *The scheme has order 2.*

Proof. This result has been experimentally demonstrated in Section 3.2 by means of the previous reasonings. □

When the exact solution is not known, the ideas above cannot be used, and some other way of checking the order must be available. This may seem strange, since we can get the order by designing problems with known exact solution, but the point here is that *the order is not observed when ψ is not regular enough*, and it is our desire to investigate the regularity of unknown exact solutions by means of numerical experiments. This justifies the introduction below of the concept of order star.

Definition 3.3.3. *Set $I \in \mathbb{N}$, $I \geq 4, I$ odd, and $N \in \mathbb{N}$, $N \geq 2$. Suppose that two embedded grids, an (I,N) grid and a $(2I-1, 2N-1)$ grid, are given, and that ψ_g and ψ_{2g} are the numerical solutions obtained for these grids by application of the odd scheme. Then the real number $D_{\text{odd}}(g,2g)$ is defined as*

$$D_{\text{odd}}(g,2g) = \max_g |\psi_g - \psi_{2g}|,$$

where the maximum is taken over all common nodes, i.e., over all the nodes of the (I,N) grid.

From Definition 3.3.3, the meaning of $D_{\text{odd}}(2g,4g)$ is evident.

Definition 3.3.4. *(*order star of the odd scheme*) Let p^* be a natural number. We say that the odd scheme has order star p^* or order* $= p^*$ if*

$$\frac{D_{\text{odd}}(g, 2g)}{D_{\text{odd}}(2g, 4g)} = 2^{p^*} \text{ asymptotically,}$$

i.e., it tends to be an equality as I and N tend to infinity.

In an analogous way, we can define the order star of the even scheme.

Definition 3.3.5. *Set $I \in \mathbb{N}$, $I \geq 4$, I even, and $N \in \mathbb{N}$, $N \geq 2$. Suppose that two embedded grids, an (I, N) grid and a $(3I - 2, 3N - 2)$ grid are given, and that ψ_g and ψ_{3g} are the numerical solutions obtained for these grids by application of the even scheme. Then the real number $D_{\text{even}}(g, 3g)$ is defined as*

$$D_{\text{even}}(g, 3g) = \max_g |\psi_g - \psi_{3g}|,$$

where the maximum is taken over all common nodes, i.e., over all the nodes of the (I, N) grid.

From Definition 3.3.5, the meaning of $D_{\text{even}}(3g, 9g)$ is evident.

Definition 3.3.6. *(*order star of the even scheme*) Let p^* be a natural number. We say that the even scheme has order star p^* or order* $= p^*$ if*

$$\frac{D_{\text{even}}(g, 3g)}{D_{\text{even}}(3g, 9g)} = 3^{p^*} \text{ asymptotically,}$$

i.e., it tends to be an equality as I and N tend to infinity.

It follows from the definitions above that the order star of the even and the odd scheme can be obtained by applying the formulas

$$p^*_{\text{odd}} = \lim_{\substack{I\uparrow\infty, N\uparrow\infty \\ I \text{ odd}}} \frac{\ln\left(\frac{D_{\text{odd}}(g,2g)}{D_{\text{odd}}(2g,4g)}\right)}{\ln 2}, \tag{3.34}$$

$$p^*_{\text{even}} = \lim_{\substack{I\uparrow\infty, N\uparrow\infty \\ I \text{ even}}} \frac{\ln\left(\frac{D_{\text{even}}(g,3g)}{D_{\text{even}}(3g,9g)}\right)}{\ln 3}, \tag{3.35}$$

Now we establish the relationship between the order and the order star. As nothing changes essentially with the parity of I, we shall restrict ourselves to the case of a general scheme that can be used either with I even or with I odd. Then the definition with factors 2 and 4 is equivalent to the definition with factors 3 and 9 and, by the same reasoning, we can freely read in the following Theorem 3.3.2 3 and 9 instead of 2 and 4. Accordingly, we shall omit the subindex "odd" or "even" that accompanies above the difference D.

Theorem 3.3.2. *Suppose that a numerical scheme has got order p. Assume moreover that*

$$D(g,2g) = \max_{g} |\psi_g - \psi| - \max_{2g} |\psi_{2g} - \psi| \quad \textit{and}$$
$$D(2g,4g) = \max_{2g} |\psi_{2g} - \psi| - \max_{4g} |\psi_{4g} - \psi|$$

asymptotically. Then the scheme has got order star equal to p.

Proof. All equalities in this proof have to be understood in an asymptotic sense.

Order p implies

$$\max_{2g} |\psi_{2g} - \psi| = \frac{\max_g |\psi_g - \psi|}{2^p} \quad \text{and} \quad \max_{4g} |\psi_{4g} - \psi| = \frac{\max_g |\psi_g - \psi|}{4^p} \tag{3.36}$$

Hence

$$D(g,2g) = \left(1 - \frac{1}{2^p}\right) \max_{g} |\psi_g - \psi| \text{ and} \tag{3.37}$$
$$D(2g,4g) = \frac{1}{2^p}\left(1 - \frac{1}{2^p}\right) \max_{g} |\psi_g - \psi|, \tag{3.38}$$

from where the desired equality

$$\frac{D(g,2g)}{D(2g,4g)} = 2^p \tag{3.39}$$

is inferred. □

Remark 3.3.3. *It is clear that Theorem 3.3.2 also holds if*

$$D(g,2g) = \max_{g} |\psi_g - \psi| + \max_{2g} |\psi_{2g} - \psi| \textit{ and}$$
$$D(2g,4g) = \max_{2g} |\psi_{2g} - \psi| + \max_{4g} |\psi_{4g} - \psi|$$

asymptotically.

Chapter 4

MATLAB® implementation

Contents

4.1. Description of the odd scheme 44

4.2. The MATLAB® command sparse 52

4.3. Defining the matrix in the code 56

4.3.1. Code for Equations *(4.1)* 58

4.3.2. Code for Equations *(4.4)* 59

4.3.3. Code for Equations *(4.2)* for $(i, n) \in \{2, \ldots, i^* - 1\} \times \{1, \ldots, N - 1\}$ 60

4.3.4. Code for Equations *(4.2)* for $(i, n) \in \{i^* + 1, \ldots, I - 1\} \times \{1, \ldots, N - 1\}$ 61

4.3.5. Code for Equations *(4.3)* 62

4.3.6. Code for Equations *(4.5)* 62

4.3.7. Code for Equations *(4.6)* 63

4.1. Description of the odd scheme

We focus on the odd scheme, as it is slightly more lengthy. In order to be clear, we rewrite the odd scheme as follows (recall that I is odd and $i^* = \frac{I+1}{2}$):

- For $(i,n) \in \{1\} \times \{1,\dots,N-1\}$,

$$\begin{aligned}\Big(-\frac{\mu_1}{k}+\frac{\overline{\alpha}_1^n}{2}+\frac{\overline{\sigma}_1^n\overline{D}_2}{2h^2}\Big)\psi_1^n+\Big(-\frac{\overline{\sigma}_1^n\overline{D}_3}{8h^2}\Big)\psi_2^n+\Big(-\frac{\overline{\sigma}_1^n\overline{D}_2}{2h^2}\Big)\psi_3^n+\Big(\frac{\overline{\sigma}_1^n\overline{D}_3}{8h^2}\Big)\psi_4^n+\\ \Big(\frac{\mu_1}{k}+\frac{\overline{\alpha}_1^{n+1}}{2}+\frac{\overline{\sigma}_1^{n+1}\overline{D}_2}{2h^2}\Big)\psi_1^{n+1}+\Big(-\frac{\overline{\sigma}_1^{n+1}\overline{D}_3}{8h^2}\Big)\psi_2^{n+1}+\Big(-\frac{\overline{\sigma}_1^{n+1}\overline{D}_2}{2h^2}\Big)\psi_3^{n+1}+\\ \Big(\frac{\overline{\sigma}_1^{n+1}\overline{D}_3}{8h^2}\Big)\psi_4^{n+1}=\frac{\overline{W}_1^n+\overline{W}_1^{n+1}}{2}. \qquad (4.1)\end{aligned}$$

- For $(i,n) \in \{2,\dots,i^*-1\} \cup \{i^*+1,\dots,I-1\} \times \{1,\dots,N-1\}$,

$$\begin{aligned}\Big(-\frac{\overline{\sigma}_i^n\overline{D}_{i-\frac{1}{2}}}{2h^2}\Big)\psi_{i-1}^n+\Big(-\frac{\mu_i}{k}+\frac{\overline{\alpha}_i^n}{2}+\frac{\overline{\sigma}_i^n\Big(\overline{D}_{i-\frac{1}{2}}+\overline{D}_{i+\frac{1}{2}}\Big)}{2h^2}\Big)\psi_i^n+\\ \Big(-\frac{\overline{\sigma}_i^n\overline{D}_{i+\frac{1}{2}}}{2h^2}\Big)\psi_{i+1}^n+\Big(-\frac{\overline{\sigma}_i^{n+1}\overline{D}_{i-\frac{1}{2}}}{2h^2}\Big)\psi_{i+1}^{n+1}+\\ \Big(\frac{\mu_i}{k}+\frac{\overline{\alpha}_i^{n+1}}{2}+\frac{\overline{\sigma}_i^{n+1}\Big(\overline{D}_{i-\frac{1}{2}}+\overline{D}_{i+\frac{1}{2}}\Big)}{2h^2}\Big)\psi_i^{n+1}+\\ \Big(-\frac{\overline{\sigma}_i^{n+1}\overline{D}_{i+\frac{1}{2}}}{2h^2}\Big)\psi_{i+1}^{n+1}=\frac{\overline{W}_i^n+\overline{W}_i^{n+1}}{2}. \qquad (4.2)\end{aligned}$$

- For $(i,n) \in \{i^*\} \times \{2,\dots,N-1\}$

$$\Big(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\Big)\psi_{i^*-1}^n+\Big(\overline{\alpha}_{i^*}^n+\frac{2\overline{\sigma}_{i^*}^n}{h^2}\Big)\psi_{i^*}^n+\Big(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\Big)\psi_{i^*+1}^n=\overline{W}_{i^*}^n. \qquad (4.3)$$

- For $(i,n) \in \{I\} \times \{1,\dots,N-1\}$,

$$\begin{aligned}\Big(\frac{\overline{\sigma}_I^n\overline{D}_{I-2}}{8h^2}\Big)\psi_{I-3}^n+\Big(-\frac{\overline{\sigma}_I^n\overline{D}_{I-1}}{2h^2}\Big)\psi_{I-2}^n+\Big(-\frac{\overline{\sigma}_I^n\overline{D}_{I-2}}{8h^2}\Big)\psi_{I-1}^n+\\ \Big(-\frac{\mu_I}{k}+\frac{\overline{\alpha}_I^n}{2}+\frac{\overline{\sigma}_I^n\overline{D}_{I-1}}{2h^2}\Big)\psi_I^n+\Big(\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-2}}{8h^2}\Big)\psi_{I-3}^{n+1}+\Big(-\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-1}}{2h^2}\Big)\psi_{I-2}^{n+1}+\\ \Big(-\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-2}}{8h^2}\Big)\psi_{I-1}^{n+1}+\Big(\frac{\mu_I}{k}+\frac{\overline{\alpha}_I^{n+1}}{2}+\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-1}}{2h^2}\Big)\psi_I^{n+1}=\frac{\overline{W}_I^n+\overline{W}_I^{n+1}}{2}. \qquad (4.4)\end{aligned}$$

- For $(i,n) \in \{i^*,\dots,I\} \times \{1\}$,

$$\psi_i^1=\overline{f}_i. \qquad (4.5)$$

- For $(i,n) \in \{1,\dots,i^*\} \times \{N\}$,

$$\psi_i^N=\overline{g}_i. \qquad (4.6)$$

Notice that, in particular, we are imposing, within Equations *(4.5)* and *(4.6)*, the numerical conditions $\psi_{i^*}^1 = f(0)$ and $\psi_{i^*}^N = g(0)$.

We seek the values of the $I \times N$ unknowns ψ_i^n for $(i,n) \in \{1,\dots,I\} \times \{1,\dots,N\}$. These values are obtained as the solution of a large linear system the coefficients of which (that is to say, the entries of the matrix and those of the second member) are defined by the numerical scheme.

We choose to pose the whole linear system, i.e., a system of order $I \times N$, even when it is true that we already know $I+1$ of these values: $I-1$ that come from the initial and final conditions *(2)* and *(3)*, that is to say, the $I-1$ values in Equations *(3.14)* and *(3.15)*, plus the two values coming from the numerical conditions *(3.16)* and *(3.17)*. The order of the system could be reduced, consequently, from $I \times N$ to $(I \times N) - (I+1)$, but the price of doing so is to spoil the structure of the scheme making it more difficult the programming task; at the same time, the computing time would not get sensibly reduced, as $I+1$ is small, we could say very small, when compared to $I \times N$.

Let us call $U \in \mathcal{M}_{I\times N}$ the matrix such that $U_{i,n} = \psi_i^n$ for $1 \le i \le I$ and $1 \le n \le N$. At this point, one must choose a way of converting U into a vector $\boldsymbol{v} \in \mathbb{R}^{I\times N}$, because finally one solves a linear system for $\boldsymbol{v}$. This is done with the help of a bijective mapping from $\{1,\dots,I\} \times \{1,\dots,N\}$ onto $\{1,\dots,I \times N\}$ which is customary to call *pointer*, and in this case we have chosen the following one:

$$\begin{array}{rcl} \{1,\dots,I\} \times \{1,\dots,N\} \subset \mathbb{N} \times \mathbb{N} & \xrightarrow{\mathrm{P}} & \{1,\dots,I \times N\} \subset \mathbb{N} \\ (i,n) & \mapsto & \mathrm{P}(i,n) = i + (n-1)I. \end{array} \tag{4.7}$$

Now, vector $\boldsymbol{v}$ containing the unknowns is defined as follows:

$$\boldsymbol{v}_{\mathrm{P(i,n)}} := U_{i,n} \text{ for } 1 \le i \le I \text{ and } 1 \le n \le N. \tag{4.8}$$

In other words,

$$\boldsymbol{v} = \begin{pmatrix} \psi_1^1 \\ \vdots \\ \psi_I^1 \\ \hline \psi_1^2 \\ \vdots \\ \psi_I^2 \\ \hline \vdots \\ \hline \psi_1^N \\ \vdots \\ \psi_I^N \end{pmatrix}. \tag{4.9}$$

As it can be guessed, the pointer plays a crucial role in the computer program, and that is why we show it as it is in our code:

```
function j = pointer(i,n,I)
% Pointer
j = i + (n - 1).*I;
end
```

Function pointer.m

Figure 4.1 shows the mesh and the linear order that pointer *(4.7)* induces in the nodes (μ_i, z_n).

In the Figure 4.1, the magenta nodes are those where the initial and final conditions (physically, the incoming flux boundary conditions) are imposed, the two red nodes are those where the numerical initial and final conditions are imposed, and the black nodes correspond to the delicate part where $\mu = 0$.

Now we introduce the concept of matrix bandwidth, which will be used in the sequel.

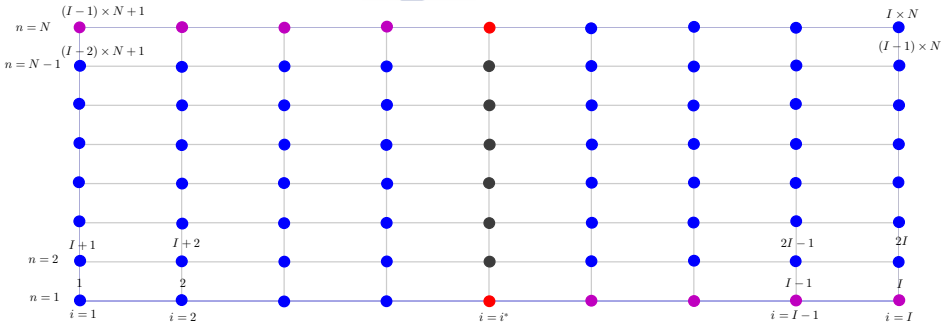


Figure 4.1: Mesh and order induced by pointer *(4.7)*.

Definition 4.1.1. Let $A = a_{(in)}$ be a real matrix of order $I \times N$. Then:

1. The *upper bandwidth* of A is the non-negative integer number $\mathfrak{b}_u(A)$ defined by

$$\mathfrak{b}_u(A) = \begin{cases} N-1 \text{ if } a_{1N} \neq 0, \\ \min\{s \in \{0,\dots,N-2\} : [a_{in} = 0 \text{ whenever } n > i+s]\} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{4.10}$$

If $\mathrm{ud}_1, \dots, \mathrm{ud}_{N-1}$ is the complete list of upper diagonals, ordered in the natural way (ud_1 the one next to the diagonal), then $s = \mathfrak{b}_u(A)$ is the minimum number such that $\mathrm{ud}_{s+1}, \dots, \mathrm{ud}_{N-1}$ are all zero vectors.
It is easy to check that

$$\mathfrak{b}_u(A) = \begin{cases} 0 \text{ if } a_{in} = 0 \quad \text{whenever} \quad i < n, \\ \max\{n-i : i < n \text{ and } a_{in} \neq 0\} \text{ otherwise.} \end{cases} \tag{4.11}$$

2. The *lower bandwidth* of A is the non-negative integer number $\mathfrak{b}_l(A)$ defined by

$$\mathfrak{b}_l(A) = \begin{cases} I-1 \text{ if } a_{I1} \neq 0, \\ \min\{s \in \{0,\dots,I-2\} : [a_{in} = 0 \text{ whenever } n < i-s]\} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{4.12}$$

If $\mathrm{ld}_1, \ldots, \mathrm{ld}_{I-1}$ is the complete list of lower diagonals, ordered in the natural way (ld_1 the one next to the diagonal), then $s = \mathfrak{b}_l(A)$ is the minimum number such that $\mathrm{ld}_{s+1}, \ldots, \mathrm{ld}_{I-1}$ are all zero vectors.
It is easy to check that

$$\mathfrak{b}_l(A) = \begin{cases} 0 \text{ if } a_{in} = 0 \quad \text{whenever} \quad i > n, \\ \max\{i - n : i > n \text{ and } a_{in} \neq 0\} \text{ otherwise.} \end{cases} \tag{4.13}$$

3. The *bandwidth* of $A \neq 0$ is the natural number $\mathfrak{b}(A)$ defined by

$$\mathfrak{b}(A) = \mathfrak{b}_u(A) + \mathfrak{b}_l(A) + 1. \tag{4.14}$$

The bandwidth of the zero matrix $A = 0$ is defined as $\mathfrak{b}(0) = 0$.
Generally speaking, one needs to know both $\mathfrak{b}_u(A)$ and $\mathfrak{b}_l(A)$ to determine which upper and lower diagonals are forming, together with the diagonal, the *band* of the matrix. Naturally, only one of these values is needed whenever $\mathfrak{b}_u(A) = \mathfrak{b}_l(A)$, for instance, when A is symmetric or has got a symmetric structure.
It is also useful to define the *balanced bandwith* of a square non-zero matrix of A as the natural number $\mathrm{b}\mathfrak{b}(A)$ defined by

$$\mathrm{b}\mathfrak{b}(A) = 2\max\{\mathfrak{b}_u(A), \mathfrak{b}_l(A)\} + 1. \tag{4.15}$$

Definition 4.1.2. The *band density* of a matrix A is given by the following formula:

$$\text{band density of } A = \frac{\text{number of non-zero entries in the band of } A}{\text{total number of entries in the band of } A}. \tag{4.16}$$

Examples:

- Let us consider the (non-square) matrix

$$A = \begin{pmatrix} \mathbf{1} & 0 & 2 & 0 \\ 0 & \mathbf{3} & 4 & 0 \\ 5 & 6 & \mathbf{7} & 8 \end{pmatrix}.$$

  Then, $\mathfrak{b}_u(A) = 2$, $\mathfrak{b}_l(A) = 2$, and $\mathfrak{b}(A) = 5$.

- Let us consider the (non-square) matrix

$$A = \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{2} & 3 & 0 \\ 0 & 4 & \mathbf{5} & 6 \end{pmatrix}.$$

  Then, $\mathfrak{b}_u(A) = 1$, $\mathfrak{b}_l(A) = 1$, and $\mathfrak{b}(A) = 3$.

- Let us consider the tridiagonal matrix

$$A = \begin{pmatrix} \mathbf{1} & 2 & 0 & 0 & 0 \\ 3 & \mathbf{4} & 5 & 0 & 0 \\ 0 & 6 & \mathbf{7} & 8 & 0 \\ 0 & 0 & 9 & \mathbf{10} & 11 \\ 0 & 0 & 0 & 12 & \mathbf{13} \end{pmatrix}.$$

Then, $\mathfrak{b}_u(A) = 1$, $\mathfrak{b}_l(A) = 1$, and $\mathfrak{b}(A) = \mathfrak{bb}(A) = 3$.

- Let us consider the matrix

$$A = \begin{pmatrix} \mathbf{1} & 2 & 3 & 0 & 0 \\ 4 & \mathbf{5} & 6 & 0 & 0 \\ 0 & 7 & \mathbf{8} & 9 & 10 \\ 0 & 0 & 10 & \mathbf{12} & 13 \\ 0 & 0 & 0 & 14 & \mathbf{15} \end{pmatrix}.$$

Then, $\mathfrak{b}_u(A) = 2$, $\mathfrak{b}_l(A) = 1$, $\mathfrak{b}(A) = 4$ and $\mathfrak{bb}(A) = 5$.

We have made the choice of the pointer, which is the choice of the order the unknowns. The other choice one must make is the order of the equations in the linear system. Let A be the matrix of this system. Typically, the order of the equations is selected so that the matrix A be "as diagonal as possible" or, in other words, so that the bandwidth of the matrix is minimized. This amounts to say that the equation related to the unknown $v_{\mathrm{P}(i,n)} = \psi_i^n$ should occupy the row $\mathrm{P}(i,n)$ within the linear system, but the point is that there are more than one equation containing ψ_i^n. As said above, all one can do is to minimize the bandwidth. The following order has been chosen according to this criterion:

- The $N-1$ Equations *(4.1)* occupy rows $\mathrm{P}(1,n)$ for $n \in \{1,\dots,N-1\}$.

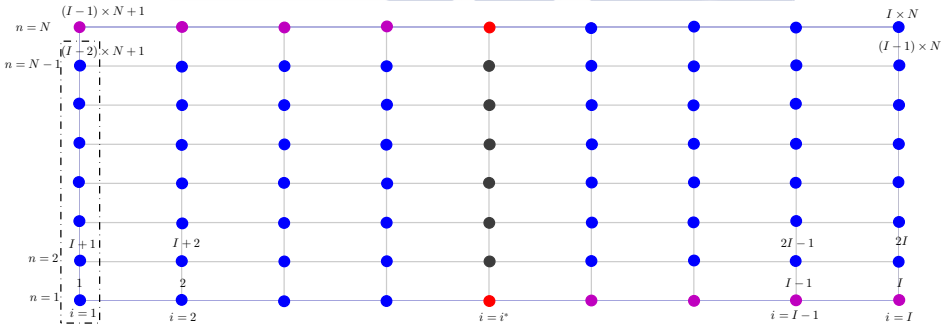


Figure 4.2: The numbers that we assign to the rows containing Equations *(4.1)* coincide with the numbers that the nodes framed in the figure have got in the global linear ordering induced by the pointer.

In this case, row $\mathrm{P}(1,n)$ has got non-zero entries only at columns $\mathrm{P}(1:4,n)$ and $\mathrm{P}(1:4,n+1)$. As the diagonal is determined by the column $\mathrm{P}(1,n)$, all these entries are located in the upper triangular part of A. Moreover, it is clear that the maximum distance between non-zero entries and the diagonal is given by

$$\mathrm{P}(4,n+1) - \mathrm{P}(1,n) = I+3. \tag{4.17}$$

- The $(i^*-2)(N-1)$ Equations *(4.2)* for indexes $(i,n) \in \{2,\dots,i^*-1\} \times \{1,\dots,N-1\}$ occupy rows $\mathrm{P}(i,n)$.

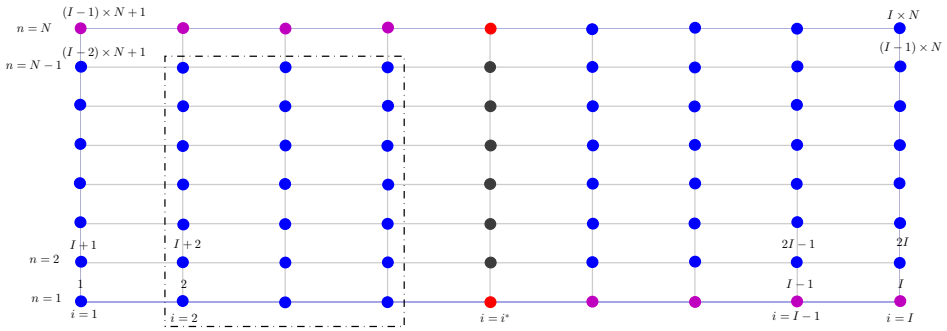


Figure 4.3: The numbers that we assign to the rows containing Equations *(4.2)* for indexes $i \in \{2, \ldots, i^* - 1\}$ coincide with the numbers that the nodes framed in the figure have got in the global linear ordering induced by the pointer.

The maximum distance between upper triangular non-zero entries and the diagonal in these rows is

$$\mathrm{P}(i+1, n+1) - \mathrm{P}(i, n) = I + 1, \tag{4.18}$$

and between lower triangular non-zero entries and the diagonal in these rows is

$$\mathrm{P}(i, n) - \mathrm{P}(i-1, n) = 1. \tag{4.19}$$

- The $(i^*-2)(N-1)$ Equations *(4.2)* for indexes $(i, n) \in \{i^*+1, \ldots, I-1\} \times \{1, \ldots, N-1\}$ occupy rows $\mathrm{P}(i, n+1)$.

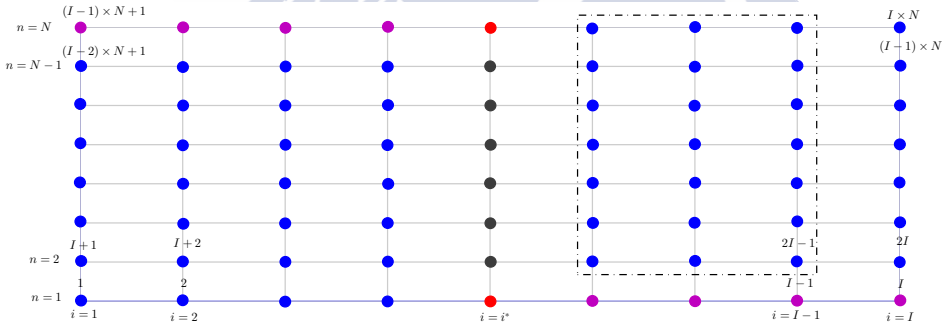


Figure 4.4: The numbers that we assign to the rows containing Equations *(4.2)* for indexes $i \in \{i^* + 1, \ldots, I - 1\}$ coincide with the numbers that the nodes framed in the figure have got in the global linear ordering induced by the pointer.

The maximum distance between upper triangular non-zero entries and the diagonal in these rows is

$$\mathrm{P}(i+1, n+1) - \mathrm{P}(i, n+1) = 1, \tag{4.20}$$

and between lower triangular non-zero entries and the diagonal in these rows is

$$\mathrm{P}(i, n+1) - \mathrm{P}(i-1, n) = I + 1. \tag{4.21}$$

- The $N - 2$ Equations *(4.3)* occupy rows $\mathrm{P}(i^*, n)$ for $n \in \{2, \ldots, N-1\}$.

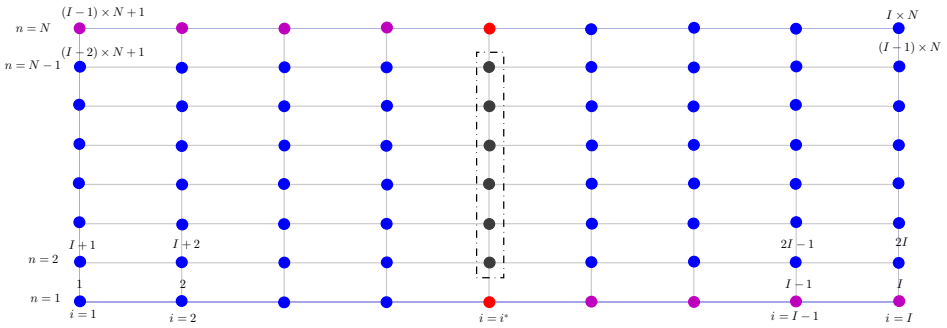


Figure 4.5: The numbers that we assign to the rows containing Equations *(4.3)* coincide with the numbers that the nodes framed in the figure have got in the global linear ordering induced by the pointer.

The maximum distance between upper triangular non-zero entries and the diagonal in these rows is

$$\mathrm{P}(i^*+1,n)-\mathrm{P}(i^*,n)=1, \tag{4.22}$$

and between lower triangular non-zero entries and the diagonal in these rows is

$$\mathrm{P}(i^*,n)-\mathrm{P}(i^*-1,n)=1. \tag{4.23}$$

- The $N-1$ Equations *(4.4)* occupy rows $\mathrm{P}(I,n+1)$ for $n\in\{1,\ldots,N-1\}$.

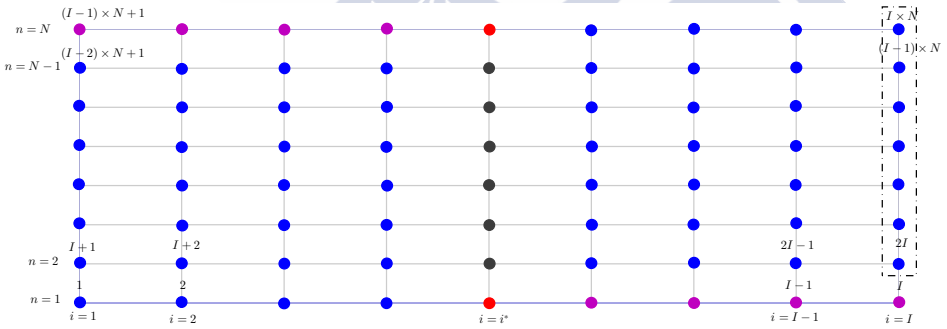


Figure 4.6: The numbers that we assign to the rows containing Equations *(4.4)* coincide with the numbers that the nodes framed in the figure have got in the global linear ordering induced by the pointer.

The maximum distance between non-zero entries and the diagonal in these rows is

$$\mathrm{P}(I,n+1)-\mathrm{P}(I-3,n)=I+3. \tag{4.24}$$

Notice that all non-zero entries are located in the lower triangular part of A.

- The i^* Equations *(4.5)* occupy rows $\mathrm{P}(i,1)$ for $i\in\{i^*,\ldots,I\}$.

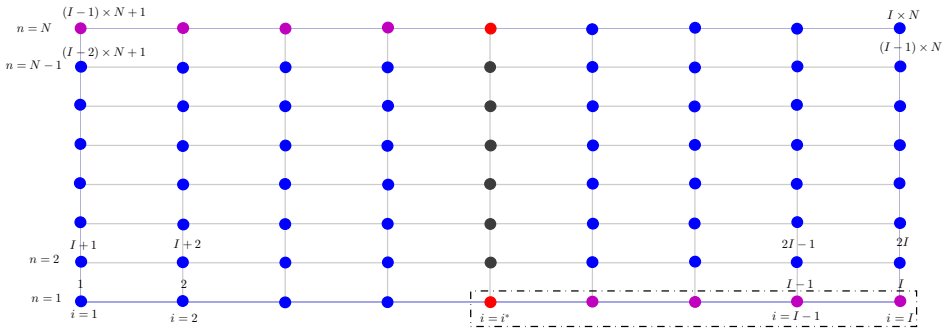


Figure 4.7: The numbers that we assign to the rows containing Equations *(4.5)* coincide with the numbers that the nodes framed in the figure have got in the global linear ordering induced by the pointer.

There is only one non-zero entry (specifically, a 1) in the diagonal of A for each of these rows.

- The i^* Equations *(4.6)* occupy rows $\mathrm{P}(i,N)$ for $i \in \{1,\ldots,i^*\}$.

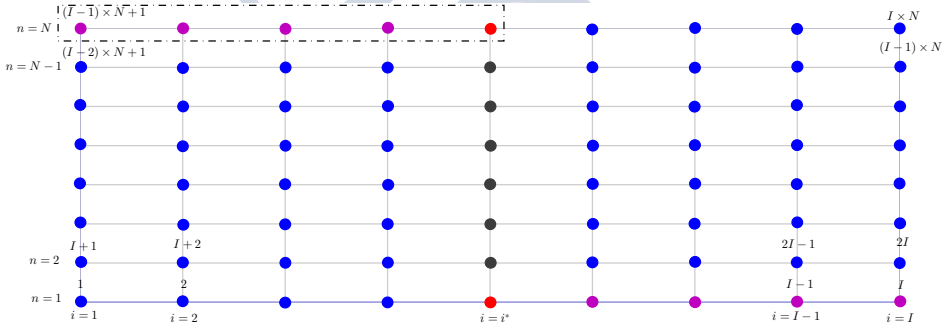


Figure 4.8: The numbers that we assign to the rows containing Equations *(4.5)* coincide with the numbers that the nodes framed in the figure have got in the global linear ordering induced by the pointer.

There is only one non-zero entry (specifically, a 1) in the diagonal of A for each of these rows.

According to Equations *(4.17)*, *(4.18)*, *(4.20)* and *(4.22)* matrix A has got upper bandwidth

$$\mathfrak{b}_u(A) = \max\{I+3, I+1, 1, 1\} = I+3, \tag{4.25}$$

and according to Equations *(4.19)*, *(4.21)*, *(4.23)* and *(4.24)*, matrix A has got lower bandwidth

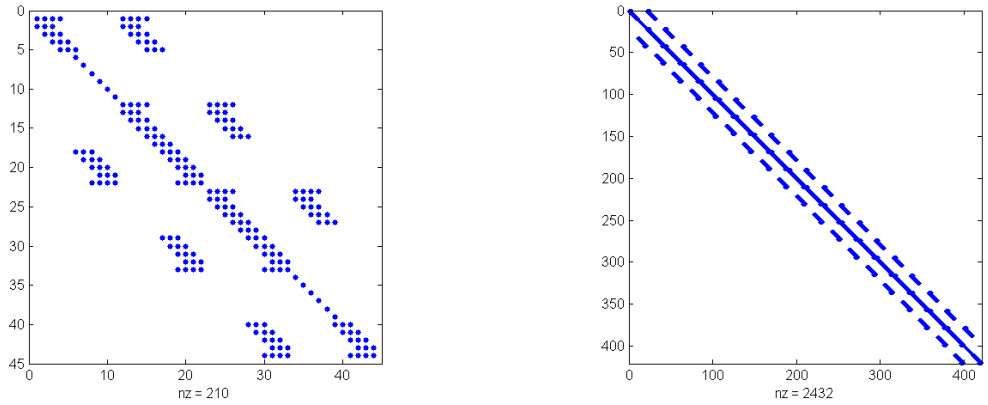
$$\mathfrak{b}_l(A) = \max\{1, I+1, 1, I+3\} = I+3, \tag{4.26}$$

which gives a total bandwidth of

$$\mathfrak{b}(A) = \mathfrak{b}_u(A) + \mathfrak{b}_l(A) + 1 = 2I+7. \tag{4.27}$$

.

Figure 4.9 shows the sparsity pattern of the matrix A for different choices of I and N, evincing that A is a banded matrix.



(a) For $I = 11$, $N = 4$. (b) For $I = 21$, $N = 20$.

Figure 4.9: Sparsity pattern of matrix A.

Once the linear system is solved, the approximation ψ_i^n can be recovered by means of Equation *(4.8)*. In our code this is done as follows:

```
V = A\sm; % Column vector living in R^dim, being
dim = I*N.
%
Mi = repmat((1:I).',1,N);
Mn = repmat(1:N,I,1);
U = V(pointer(Mi,Mn,I)); % U(i,n)= V(j) when j =
pointer(i,n,I).
```

Part of the code that solves the linear system and recovers from $\boldsymbol{v} = V$ the unknowns in matrix form $U = U$.

The code cannot be understood without knowing how the command `sparse` works.

4.2. The MATLAB® command `sparse`

This command allows defining sparse matrices in the logical way, which is disregarding all zero entries. When a matrix is sparse, it is a big mistake to treat it as a full one. Indeed, if the matrix is defined by means of the `sparse` command, both the memory requirements and the computing time (for solving an associated linear system) will be drastically reduced.

The `sparse` command works as follows: let us suppose that, in

$$\texttt{A = sparse(r,c,ent,dim1,dim2)}, \tag{4.28}$$

`r` (for "rows"), `c` (for "columns") and `ent` (for "entries") are matrices of the same order $M \times N$, and that this order is not larger than `dim1`$\times$`dim2`, typically much smaller in fact, as we are thinking of sparse matrices. Anyway, what follows has got sense if $1 \leq M \leq$`dim1` and

$1 \leq N \leq$dim2. Notice that `r, c` and `ent` need not be matrices in a strict sense, but they can be scalar or vector quantities as well; see examples below.

Then, Equation *(4.28)* will define the matrix `A` characterized as follows:

$$\mathtt{A} \in \mathcal{M}_{\mathtt{dim1}\times\mathtt{dim2}}, \tag{4.29}$$

$$\mathtt{A}_{i,j} = 0 \text{ if } (i,j) \notin \{(r(k,l), c(k,l))\}_{1\leq k\leq M,\, 1\leq l\leq N}, \tag{4.30}$$

$$\mathtt{A}_{\mathtt{rows}(i,j),\mathtt{columns}(i,j)} = \sum_{\{(k,l):\, (\mathtt{r}(k,l),\mathtt{c}(k,l))=(r(i,j),c(i,j))\}} \mathtt{ent}(k,l). \tag{4.31}$$

Often, pairs `r(i,j),c(i,j)` are not repeated. In this case, which will be always ours when coding the scheme, Equation *(4.31)* takes the simpler form

$$\mathtt{A}_{\mathtt{rows}(i,j),\mathtt{columns}(i,j)} = \mathtt{ent}(k,l). \tag{4.32}$$

The command `full(A)` writes down the full matrix.

- **First example:**

```
>> A = sparse(1,3,-4,4,4,4)
A =
  (1,3)    -4
>> full(A)
ans =
  0    0    -4    0
  0    0     0    0
  0    0     0    0
  0    0     0    0
```

- **Second example:**

```
>> A = sparse([1,2],[3,1],[-4,6],4,4)
A =
  (2,1)     6
  (1,3)    -4
>> full(A)
ans =
  0    0    -4    0
  6    0     0    0
  0    0     0    0
  0    0     0    0
```

- **Third example:**

```
>> A = sparse([1,2;3,2;4,4],[3,1;1,3;2,1],[-4,6;10,11;9,3],4,4)
A =
```

```
  (2,1)    6
  (3,1)   10
  (4,1)    3
  (4,2)    9
  (1,3)   -4
  (2,3)   11
>> full(A)
ans =
   0    0   -4    0
   6    0   11    0
  10    0    0    0
   3    9    0    0
```

- **Fourth example:**

```
>> A = sparse([1,2;3,2;4,4],[3,1;1,1;2,1],[-4,6;10,11;9,3],4,4)
A =
  (2,1)   17
  (3,1)   10
  (4,1)    3
  (4,2)    9
  (1,3)   -4
>> full(A)
ans =
   0    0   -4    0
  17    0    0    0
  10    0    0    0
   3    9    0    0
```

What is happening here? Notice that we are providing two values, 6 and 11, for the same position (2, 1). In this case, as antipated by Equation *(4.29)*, MATLAB® adds these two values.

- **Fifth example** (construction of a symmetric tridiagonal matrix of order 4×4), with diagonal vector `dv`, and upper diagonal `udv`):

```
>> dv = 2.*ones(1,4)
dv=
  2   2   2   2
>> dA = sparse(1:4,1:4,dv,4,4)
  (1,1)    2
  (2,2)    2
  (3,3)    2
  (4,4)    2
>> full(dA)
ans =
```

```
  2    0    0    0
  0    2    0    0
  0    0    2    0
  0    0    0    2
>> udv = -1.*ones(1,3)
udv =
  -1    -1    -1
>> udA = sparse(1:3,2:4,udv,4,4)
udA =
  (1,2)    -1
  (2,3)    -1
  (3,4)    -1
>> full(udA)
ans =
  0    -1     0     0
  0     0    -1     0
  0     0     0    -1
  0     0     0     0
>> A = dA + udA + udA.'
A =
  (1,1)     2
  (2,1)    -1
  (1,2)    -1
  (2,2)     2
  (3,2)    -1
  (2,3)    -1
  (3,3)     2
  (4,3)    -1
  (3,4)    -1
  (4,4)     2
>> full(A)
ans =
   2    -1     0     0
  -1     2    -1     0
   0    -1     2    -1
   0     0    -1     2
```

In brief, one could write:

```
>> dv = 2.*ones(1,4); udv = -1.*ones(1,3);
>> dA = sparse(1:4,1:4,dv,4,4); udA = sparse(1:3,2:4,udv,4,4);
>>  A = dA + udA + udA.';
```

Of course, one can also write, with the same effect:

```
>> dv = 2.*ones(1,4); udv = -1.*ones(1,3);
>> A = sparse([1:4,1:3,2:4],[1:4,2:4,1:3],[dv,udv,udv],4,4);
```

4.3. Defining the matrix in the code

In this Section we are going to explain a good way to define the matrix of the system by means of sparse command. The reason is that the way in which the matrix is defined has got a very strong effect in the computing time. The basic idea is to exploit the vector abilities of MATLAB®.

The matrix can be easily identified by looking at the scheme description in Section 4.1. As said before, we focus on the odd scheme.

In order to quantify the degree of sparsity, let us note that the number of total matrix entries is $(I \times N)^2 = (IN)^2$, while the total number of non-zero entries is only $6(I \times N) - 5I + N - 3 = 6IN - 5I + N - 3$. It is evident that the percentage of non-zero entries will be asymptotically equal to $\frac{6\times 100}{I\times N}$ (see Table 4.3.1).

I	N	$(IN)^2$	$6IN - 5I + N - 3$	non-zero entries
11	10	12100	612	5.06%
11	100	1210000	6642	5.49×10^{-1}%
101	100	102010000	60192	5.90×10^{-2}%
101	1000	10201000000	606492	5.94×10^{-3}%
1001	1000	1002001000000	6001992	5.99×10^{-4}%

Table 4.3.1: Table showing sparsity.

The *band density* is another figure of matter, because it is used by MATLAB® to decide which method will be used for solving the linear system when (as it is our case) one has got a banded matrix. This quantity is defined as the quotient between the number of non-zero entries and the total number of entries in the band of the matrix. The first of these two numbers has already been stated, while the total number of entries in the band is given by

$$\begin{aligned}(I \times N) + 2[(I \times N) - 1] + 2[(I \times N) - 2] + \cdots + 2[(I \times N) - (I + 2)] - \\ 2[(I \times N) - (I + 3)] = [2(I + 3) + 1](I \times N) - (I + 3)(N + 4)\end{aligned} \tag{4.33}$$

because both the upper and the lower bandwidth are equal to $I + 3$. Table 4.3.2 shows values of the band density for several choices of I and N.

I	N	$[2(I + 3) + 1]IN - (I + 3)(I + 4)$	$6IN - 5I + N - 3$	band density
11	10	2980	612	0.2054
11	100	31690	6642	0.2096
101	100	2099980	60192	0.0287
101	1000	21098080	606492	0.0287
1001	1000	2009999980	6001992	0.0030

Table 4.3.2: Table showing band density.

The number $6IN-5I+N-3$ used above can be easily computed taking in account that:

- $8(N-1)$ non-zero entries come from Equation *(4.1)*.
- $6(I-3)(N-1)$ non-zero entries come from Equation *(4.2)*.
- $3(N-2)$ non-zero entries come from Equation *(4.3)*.
- $8(N-1)$ non-zero entries come from Equation *(4.4)*.
- $I+1$ non-zero entries come from Equation *(4.5)–(4.6)*.

To get the result, we simply do the sum:

$$8(N-1)+6(I-3)(N-1)+3(N-2)+8(N-1)+I+1=6IN-5I+N-3. \quad (4.34)$$

In the next Subsections we will explain how we have defined the matrix in our code. In order to understand it, it is highly recommended to conduct some tests and experiments with both the repmat and kron commands in simple cases. The following notations will be used within the code:

- `dim` $= I \times N$.
- `A` will be, at the end of the process, the matrix of the linear system. Different types of equations are progressively assembled for constructing A.
- `sm`: second member.

4.3.1. Code for Equations *(4.1)*

As said before, we have decided that the $N-1$ equations contained in Equation *(4.1)* occupy in our linear system the rows $\mathrm{P}(1,n)$ for $n \in \{1,\ldots,N-1\}$.

The code is as follows:

```
indexes = pointer(1,(1:N-1).',I); % Indexes of the equations in the global
ordering
rows = repmat(indexes,1,8); % Because there will be 8 non-zero entries at
every row
Mi = repmat([1:4,1:4],N-1,1); % myu-nodes with indexes 1, 2, 3 and 4 are
involved
Mn = [repmat((1:N-1).',1,4), repmat((2:N).',1,4)]; % 4 coefficients at
time n, plus 4 at time n+1
columns = pointer(Mi,Mn,I);
mu1 =myu(1); mu2 = myu(2); mu3 = myu(3);
D2 = D(mu2); D3 = D(mu3);
z1 = z(1:N-1).'; z2 = z(2:N).';
alph1 = alpha(mu1,z1); alph2 = alpha(mu1,z2);
sig1 = sigma(mu1,z1); sig2 = sigma(mu1,z2);
C1 = -(mu1./k) + alph1./2 + sig1.*D2./(2*h^2);
C2 = -sig1.*D3./(8*h^2);
C3 = -sig1.*D2./(2*h^2);
C4 = -C2;
C5 = (mu1./k) + alph2./2 + sig2.*D2./(2*h^2);
C6 = -sig2.*D3./(8*h^2);
C7 = -sig2.*D2./(2*h^2);
C8 = -C6;
COEFS = [C1,C2,C3,C4,C5,C6,C7,C8];
A = sparse(rows,columns,COEFS,dim, dim);
sm(indexes) = (W(mu1,z1) + W(mu1,z2))./2;
```

Part of the code defining the rows $\mathrm{P}(1,n)$ for $n \in \{1,\ldots,N-1\}$.

4.3.2. Code for Equations *(4.4)*

As said before, the $N-1$ equations contained in Equation *(4.4)* occupy in our linear system the rows $\mathrm{P}(I, n+1)$ for $n \in \{1, \ldots, N-1\}$.

The code is as follows:

```
indexes = pointer(I,(2:N).',I); % Indexes of the equations in the global
ordering.
rows = repmat(indexes,1,8); % Because there will be 8 non-zero entries at
every row
Mi = repmat([I-3:I,I-3:I],N-1,1); % myu-nodes with indexes I-3, I-2,I-1
and I are involved
Mn = [repmat((1:N-1).',1,4), repmat((2:N).',1,4)]; % 4 coefficients at
time n, plus 4 at time n+1
columns = pointer(Mi,Mn,I);
muIm2 = myu(I-2); muIm1 = myu(I-1); muI = myu(I);
DIm2 = D(muIm2); DIm1 = D(muIm1);
z1 = z(1:N-1).'; z2 = z(2:N).';
alph1 = alpha(muI,z1); alph2 = alpha(muI,z2);
sig1 = sigma(muI,z1); sig2 = sigma(muI,z2);
C1 = sig1.*DIm2./(8*h^2);
C2 = -sig1.*DIm1./(2*h^2);
C3 = -C1;
C4 = -(muI/k) + alph1./2 + sig1.*DIm1./(2*h^2);
C5 = sig2.*DIm2./(8*h^2);
C6 = -sig2.*DIm1./(2*h^2);
C7 = -C5;
C8 = (muI/k) + alph2./2 + sig2.*DIm1./(2*h^2);
COEFS = [C1,C2,C3,C4,C5,C6,C7,C8];
A = A + sparse(rows,columns,COEFS,dim, dim);
sm(indexes) = (W(muI,z1) + W(muI,z2))./2;
```

Part of the code adding the rows $\mathrm{P}(I, n+1)$ for $n \in \{1, \ldots, N-1\}$ to the already defined rows.

4.3.3. Code for Equations *(4.2)* for $(i,n) \in \{2,\dots,i^*-1\} \times \{1,\dots,N-1\}$

As said before, the $(i^*-2)(N-1)$ equations which are contained in Equation *(4.2)* for $(i,n) \in \{2,\dots,i^*-1\} \times \{1,\dots,N-1\}$ occupy in our linear system the rows $\mathrm{P}(i,n)$.

The code is as follows:

```
istar = (I + 1)/2;
ivector = 2:istar-1;
Mi = repmat(ivector.',1,N-1);
Mn = repmat(1:N-1,istar-2,1);
indexes = pointer(Mi,Mn,I); % Indexes of the equations in the global ordering.
vi = kron(ivector.',ones(N-1,1));
rMi = repmat(vi,1,6);
rMn = repmat(repmat((1:N-1).',1,6),istar-2,1);
rows = pointer(rMi,rMn,I);
cMi = repmat([vi-1,vi,vi+1],1,2);
cMn = repmat([repmat((1:N-1).',1,3),repmat((2:N).',1,3)],istar-2,1);
columns = pointer(cMi,cMn,I);
vn = repmat((1:N-1).',istar-2,1);
vmu = myu(vi); vz1 = z(vn); vz2 = z(vn + 1);
alph1 = alpha(vmu,vz1); alph2 = alpha(vmu,vz2);
sig1 = sigma(vmu,vz1); sig2 = sigma(vmu,vz2);
D1 = D(vmu - h/2); D2 = D(vmu + h/2);
C1 = -sig1.*D1./(2*h^2);
C2 = -(vmu./k) + alph1./2 + sig1.*(D1 + D2)./(2*h^2);
C3 = -sig1.*D2./(2*h^2);
C4 = -sig2.*D1./(2*h^2);
C5 = (vmu./k) + alph2./2 + sig2.*(D1 + D2)./(2*h^2);
C6 = -sig2.*D2./(2*h^2);
COEFS = [C1,C2,C3,C4,C5,C6];
A = A + sparse(rows,columns,COEFS,dim, dim);
mui = myu(Mi);zn1 = z(Mn); zn2 = z(Mn + 1);
sm(indexes) = (W(mui,z1) + W(mui,z2))./2;
```

Part of the code adding the rows $\mathrm{P}(i,n)$ for $(i,n) \in \{2,\dots,i^*-1\} \times \{1,\dots,N-1\}$ to the already defined rows.

4.3.4. Code for Equations *(4.2)* for $(i,n) \in \{i^*+1,\dots,I-1\} \times \{1,\dots,N-1\}$

As said before, the $(i^*-2)(N-1)$ equations which are contained in Equation *(4.2)* for $(i,n) \in \{i^*+1,\dots,I-1\} \times \{1,\dots,N-1\}$ occupy in our linear system the rows $\mathrm{P}(i,n+1)$.

The code is as follows:

```
ivector = istar+1:I-1;
Mi = repmat(ivector.',1,N-1);
Mn = repmat(2:N,istar-2,1);
indexes = pointer(Mi,Mn,I); % Indexes of the equations in the global ordering.
vi = kron(ivector.',ones(N-1,1));
rMi = repmat(vi,1,6);
rMn = repmat(repmat((1:N-1).',1,6),istar-2,1);
rows = pointer(rMi,rMn,I);
cMi = repmat([vi-1,vi,vi+1],1,2);
cMn=repmat([repmat((1:N-1).',1,3),repmat((2:N).',1,3)],istar-2,1);
columns = pointer(cMi,cMn,I);
vn = repmat((1:N-1).',istar-2,1);
vmu = myu(vi); vz1 = z(vn); vz2 = z(vn + 1);
alph1 = alpha(vmu,vz1); alph2 = alpha(vmu,vz2);
sig1 = sigma(vmu,vz1); sig2 = sigma(vmu,vz2);
D1 = D(vmu - h/2); D2 = D(vmu + h/2);
C1 = -sig1.*D1./(2*h^2);
C2 = -(vmu./k) + alph1./2 + sig1.*(D1 + D2)./(2*h^2);
C3 = -sig1.*D2./(2*h^2);
C4 = -sig2.*D1./(2*h^2);
C5 = (vmu./k) + alph2./2 + sig2.*(D1 + D2)./(2*h^2);
C6 = -sig2.*D2./(2*h^2);
COEFS = [C1,C2,C3,C4,C5,C6];
A = A + sparse(rows,columns,COEFS,dim,dim);
mui = myu(Mi);zn1 = z(Mn - 1); zn2 = z(Mn);
sm(indexes) = (W(mui,z1) + W(mui,z2))./2;
```

Part of the code adding the rows $\mathrm{P}(i,n+1)$ for $(i,n) \in \{i^*+1,\dots,I-1\} \times \{1,\dots,N-1\}$ to the already defined rows.

4.3.5. Code for Equations *(4.3)*

As said before, the $N-2$ equations contained in Equation *(4.3)* occupy in our linear system the rows $\mathrm{P}(i^*, n)$ for $n \in \{2, \ldots, N-1\}$.

The code is as follows:

```
indexes = pointer(istar,(2:N-1).',I);% Indexes of the equations in the
global ordering.
rows = repmat(indexes,1,3);
Mi = repmat(istar-1:istar+1,N-2,1);
Mn = repmat((2:N-1).',1,3);
columns = pointer(Mi,Mn,I);
mui = myu(istar); z1 = z(2:N-1).';
alph = alpha(mui,z1); sig = sigma(mui,z1);
C1 = -sig/(h^2);
C2 = alph + 2.*sig./(h^2);
% C3 = C1;
COEFS =[C1,C2,C1];
A = A + sparse(rows,columns,COEFS,dim,dim);
sm(indexes) = W(mui,z1);
```

Part of the code adding the rows $\mathrm{P}(i^*, n)$ for $n \in \{2, \ldots, N-1\}$ to the already defined rows.

4.3.6. Code for Equations *(4.5)*

As said before, the i^* equations contained in Equation *(4.5)* occupy in our linear system the rows $\mathrm{P}(i, 1)$ for $i \in \{i^*, \ldots, I\}$.

The code is as follows:

```
vi = istar:I;
indexes = pointer(vi.',1,I);
mu1 = myu(vi).';
% rows = columns = indexes
unos = ones(1,istar);
A = A + sparse(indexes,indexes,unos,dim,dim);
sm(indexes) = f(mu1);
```

Part of the code adding the rows $\mathrm{P}(i, 1)$ for $i \in \{i^*, \ldots, I\}$ to the already defined rows.

4.3.7. Code for Equations *(4.6)*

As said before, the i^* equations contained in Equation *(4.6)* occupy in our linear system the rows $\mathrm{P}(i, N)$ for $i \in \{1, \ldots, i^*\}$.

The code is as follows:

```
vi = 1:istar;
indexes = pointer(vi.',1,I);
mu1 = myu(vi).';
% rows = columns = indexes
unos = ones(1,istar);
A = A + sparse(indexes,indexes,unos,dim,dim);
sm(indexes) = g(mu1);
```

Part of the code adding the rows $\mathrm{P}(i, N)$ for $i \in \{1, \ldots, i^*\}$ to the already defined rows.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 5

Iterative method

Contents

5.1. Introduction 66

5.2. Preliminaries 66

5.3. Description of the iterative algorithm 67

5.1. Introduction

In this Chapter we study an iterative method, which is slightly different than the ones used in [21] and [32] to solve the problem *(1)–(3)*. In the next Chapter numerical results are presented obtained with this method and these results are compared with those obtained in Chapter 3.2.

As in Introduction, we consider the model problem defined on

$$Q = [-1, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}] \tag{5.1}$$

by the forward-backward diffusion PDE

$$\mu \frac{\partial \psi}{\partial z} + \alpha\psi - \sigma \frac{\partial}{\partial \mu}\left[D(\mu)\frac{\partial \psi}{\partial \mu}\right] = W \text{ for } (\mu, z) \in Q, \tag{5.2}$$

where $D(\mu) = 1 - \mu^2$; $\alpha \geq 0$, $\sigma > 0$ and W are known functions of (μ, z), together with the incoming flux boundary conditions

$$\psi(\mu, Z_{\text{ini}}) = f(\mu) \quad \text{for } \mu \in (0, 1], \tag{5.3}$$
$$\psi(\mu, Z_{\text{fin}}) = g(\mu) \quad \text{for } \mu \in [-1, 0). \tag{5.4}$$

In respect of hypotheses on the data functions, we shall assume that

$$\alpha,\ \sigma,\ W \text{ are continuous on } Q, \text{ and } \alpha \geq 0,\ \sigma > 0, \tag{5.5}$$
$$f \text{ is continuous on } [0, 1],\ g \text{ is continuous on } [-1, 0]. \tag{5.6}$$

The Chapter focuses on showing that *a direct approach is easily advantageous over an iterative approach* when solving numerically the problem *(5.2)–(5.4)*. Details given in the sequel will make this assertion definite.

The scope of this assertion goes beyond the examples studied, and can be applied to other forward-backward diffusion problems appearing in the literature, such as the boundary value problem investigated in [14] or the forward-backward heat equation as it stands in [33] when $a = a(x)$.

5.2. Preliminaries

First of all we recall some comments in Section 3.1: Consider the following subsets of Q: $Q_- = [-1, 0) \times [Z_{\text{ini}}, Z_{\text{fin}}]$, $Q_+ = (0, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}]$ and $Q_0 = \{0\} \times [Z_{\text{ini}}, Z_{\text{fin}}]$ (see Figure 5.1).

Let $\{(\mu_i, z_n) : i \in \{1, \ldots, I\}, \{1, \ldots, N\}\}$ be a mesh of Q obtained as the Cartesian product of two uniform meshes of $[-1, 1]$ and $[Z_{\text{ini}}, Z_{\text{fin}}]$, and think of using a finite scheme over this mesh.

In Chapter 4 we asserted that in order to solve the problem *(5.2)–(5.4)* one must think (the quoted text below defines, by the way, what we mean by *iterative* and *direct* approach)

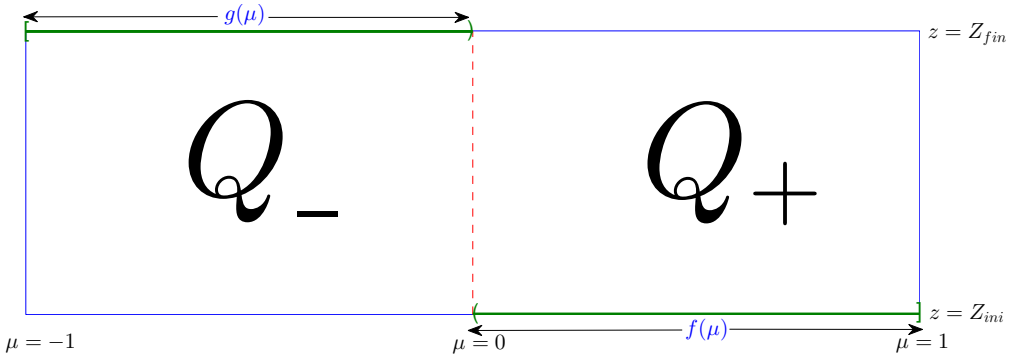


Figure 5.1: Domain $Q = Q_- \cup Q_0 \cup Q_+$, with indication of the incoming flux boundary conditions. The dotted vertical line represents the subset Q_0.

- either of using an iterative process starting with initial guess of the solution on Q_0,
- or of using a global scheme like that used for solving the two-dimensional Poisson equation, where the approximations at all mesh points (μ_i, z_n) are simultaneously obtained as the solution of a single large (but sparse) linear system.

After these comments, only the second line of action (i.e., the direct approach) was followed in Chapters 3 and 4. In view of the fact that some scientists have manifested through personal communications that the first one should require less computing time and after checking that all methods published before were of iterative type, we find it necessary to carry out a series of numerical experiments, a selection of which is collected in the Chapter 6 where it is allowed us to compare iterative and direct methods.

5.3. Description of the iterative algorithm

The aim of the iterative algorithm is to compute the solution of the same linear system given by the equations *(4.1)–(4.6)* in Section 4.1 by decomposing Q into Q_- and Q_+ and observing that it is possible to use a marching strategy in the variable z by working separately on each of these subdomains. Clearly, one advances when solving on Q_+, but goes back when solving on Q_- (see Figure 5.1).

In order to start this procedure, an initial guess of the solution on the points of the mesh belonging to Q_0 is required. The set of these values is updated after performing one forward march on Q_+ and one backward march on Q_- until convergence is achieved or a given number of iterations has been carried out.

In application to particular cases of our problem, and also to other forward-backward PDEs, proofs of convergence for iterative algorithms based on the same guidelines can be found (see [14], [18] and [32], [33]).

The detailed description reads as follows (every time we say "on Q_0" we actually mean "on the relative interior of Q_0", that is, on $\{0\} \times (Z_{\text{ini}}, Z_{\text{fin}})$):

STEP 0. **The seed:**
Provide the algorithm with an initial guess of the values of the solution on Q_0:

$$\left(\psi_{i^*}^n\right)^{[0]} \text{ for } n \in \{2, \ldots, N-1\}. \tag{5.7}$$

This set of values is usually called the *seed* of the iterative algorithm. Figure 5.2 shows the points of the mesh where the seed is given.

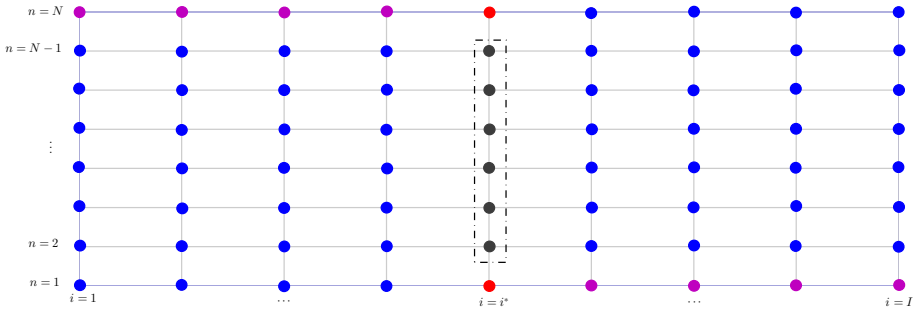


Figure 5.2: The points where the seed is required have been boxed.

A good choice of the seed, which amounts to say choosing values close to the exact solution of the discrete problem, will make the iterative process converge in fewer iterations. Maybe the simplest reasonable possibility lies in the idea of joining the pairs $(f(0), Z_{\text{ini}})$ and $(g(0), Z_{\text{fin}})$ by means of a straight line:

$$\left(\psi_{i^*}^n\right)^{[0]} = f(0) + \left[g(0) - f(0)\right]\frac{(n-1)}{N-1} \text{ for } n \in \{2, \ldots, N-1\}. \tag{5.8}$$

However, a better seed can be obtained as follows: firstly, solve the problem with the direct method using a coarse grid (in order to spend only a negligible amount of time), and then take the restriction to Q_0 of this numerical solution as the basis to calculate the seed in Equation *(5.7)* by means of linear interpolation. Neither spline interpolation nor piecewise cubic Hermite interpolation, the other two possibilities offered by our MATLAB® version, has outperformed the linear one in any of our numerical experiments.

STEP 1. **Computation of values on Q_+ at iteration $q+1$:**
For a fixed number $q \in \mathbb{N} \cup \{0\}$, assume that

$$\left(\psi_{i^*}^n\right)^{[q]} \text{ for } n \in \{2, \ldots, N-1\} \tag{5.9}$$

are known, and obtain all values on Q_+ at iteration $q+1$, i.e.,

$$\left(\psi_i^n\right)^{[q+1]} \text{ for } (i,n) \in \{i^*+1, \ldots, I\} \times \{1, \ldots, N\}, \tag{5.10}$$

by solving Equations (*(4.2)*, for $i \in \{i^*+1, \ldots, I-1\}$), *(4.4)* and *(4.5)*, with the aid of the boundary values *(5.9)*.
This is done by means of a one-step forward marching procedure starting from the known values at $z = Z_{\text{ini}}$ given by equation *(4.5)*.

STEP 2. **Computation of values on Q_- at iteration $q+1$:**
Obtain all values on Q_- at the same iteration $q+1$, i.e.,

$$\left(\psi_i^n\right)^{[q+1]} \text{ for } (i,n) \in \{1, \ldots, i^*-1\} \times \{1, \ldots, N\}, \tag{5.11}$$

by solving Equations *(4.1)*, (*(4.2)*, for $i \in \{2, \dots, i^* - 1\}$), and *(4.6)*, with the aid of the same boundary values *(5.9)*.
This is done by means of a one-step backward marching procedure starting from the known values at $z = Z_{\text{fin}}$ given by equation *(4.6)*.

STEP 3. **Update (computation of values on Q_0 at iteration $q + 1$:)**
Update of the values on Q_0 can be done by using Equation *(4.3)*. That is to say, the updated values

$$\left(\psi_{i^*}^n\right)^{[q+1]} \text{ for } n \in \{2, \dots, N-1\} \tag{5.12}$$

can follow from

$$\left(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\right)\left(\psi_{i^*-1}^n\right)^{[q+1]} + \left(\overline{\alpha}_{i^*}^n + \frac{2\overline{\sigma}_{i^*}^n}{h^2}\right)\left(\psi_{i^*}^n\right)^{[q+1]} + \left(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\right)\left(\psi_{i^*+1}^n\right)^{[q+1]} = \overline{W}_{i^*}^n. \tag{5.13}$$

However, it is possible to accelerate convergence by over-relaxing the update, in the way we explain now: firstly, one computes

$$\left(\tilde{\psi}_{i^*}^n\right)^{[q+1]} \text{ for } n \in \{2, \dots, N-1\} \tag{5.14}$$

from

$$\left(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\right)\left(\psi_{i^*-1}^n\right)^{[q+1]} + \left(\overline{\alpha}_{i^*}^n + \frac{2\overline{\sigma}_{i^*}^n}{h^2}\right)\left(\tilde{\psi}_{i^*}^n\right)^{[q+1]} + \left(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\right)\left(\psi_{i^*+1}^n\right)^{[q+1]} = \overline{W}_{i^*}^n, \tag{5.15}$$

and, finally, the updated values are computed from

$$\left(\psi_{i^*}^n\right)^{[q+1]} = \omega\left(\tilde{\psi}_{i^*}^n\right)^{[q+1]} + (1-\omega)\left(\psi_{i^*}^n\right)^{[q]} \tag{5.16}$$

for $n \in \{2, \dots, N-1\}$, where $\omega \in \mathbb{R}$ is a given relaxation parameter. Notice that $\omega = 1$ means that no relaxation is being deployed. The value ω cannot be 0, otherwise no update is taking place, but being different from 0 is not guarantee of convergence. Among those values of ω offering convergence, the optimal one should be employed, but as of today the issue of finding a closed expression for this optimum (or for an estimation of it) has not been investigated; accordingly, the values of ω used in Chaper 6 when reporting the numerical results have been found by performing several trials, looking for values that reduce the number of iterations in a significant way with respect to the case $\omega = 1$.
The idea of relaxing the update in forward-backward diffusion problems has been already used in [32] and [33].

STEP 4. **Checking convergence:**
Let us define the vectors $\boldsymbol{u}^{[q]}$, $\boldsymbol{u}^{[q+1]} \in \mathbb{R}^{N-2}$ as follows:

$$u_j^{[q]} = \left(\psi_{i^*}^{j+1}\right)^{[q]}, \quad u_j^{[q+1]} = \left(\psi_{i^*}^{j+1}\right)^{[q+1]} \tag{5.17}$$

for $j \in \{1, \dots, N-2\}$.
Given a small real number $\varepsilon > 0$, we say that the algorithm has converged with ε-tolerance if

$$\frac{\|\mathbf{u}^{[q+1]} - \mathbf{u}^{[q]}\|}{1 + \|\mathbf{u}^{[q+1]}\|} \leq \varepsilon, \tag{5.18}$$

where $\|\cdot\|$ stands for the Euclidean norm in $\mathbb{R}^{N-2}$. The quotient in Equation *(5.18)* will be referred to as the *residual*.

In case that condition *(5.18)* is not satisfied, one sets $q = q + 1$ and goes back to STEP 1. If on the contrary condition *(5.18)* holds, then the process finishes and the last computed values are taken as the ultimate numerical solution:

$$\psi_i^n = \left(\psi_i^n\right)^{[q+1]} \quad \text{for} \quad (i,n) \in \{1,\dots,I\} \times \{1,\dots,N\}. \tag{5.19}$$

Notice that both the solution *(5.19)* and the number of iterations needed for getting it depend on the seed, on the relaxation parameter ω, on the tolerance parameter ε and on the norm used in the stopping criterium *(5.18)*.

In the sequel, an *iteration* will be the process which consists of performing steps 1, 2, 3 and 4.

Chapter 6

Direct method versus Iterative method

Contents

6.1. Prefatory comments 72

6.2. Test #1 72

6.2.1. Results obtained with the iterative method 73

6.2.2. Results obtained with the direct method 75

6.3. Test #2 76

6.3.1. Results obtained with the iterative method 77

6.3.2. Results obtained with the direct method 77

6.4. Some additional comments 78

In this Chapter we will report and compare numerical results obtained with both direct and iterative methods, which experimentally show that the direct method outperforms the iterative one unless a very fine mesh is required.

6.1. Prefatory comments

In this and next Section, $Z_{\text{ini}} = 0$, $Z_{\text{fin}} = 1$ and $E_{\text{abs}}(Q) = \max_Q |\psi_{\text{grid}} - \psi|$, where ψ_{grid} is representing the approximate solution and the maximum is taken over the set of all nodes. Computations have been performed by running MATLAB®, R2015a, on a personal computer with an INTEL® Core i7-4790 @ 3.60GHz processor.

MATLAB® automatically chooses, via the single character "\", direct methods for solving the linear systems involved: LAPACK for the linear systems of the iterative method, and UMFPACK for the only linear system of the direct method. More information about these routines can be found in the reference [5].

All matrices involved are sparse and banded. In the case of the direct method, we have vectorized, in the code, the definition of the matrix in order to avoid loops, and as it is natural we have imposed an ordering of the equations and unknowns leading to small bandwidth. We remark that the matrices needed in the iterative algorithm are much easier to deal with, and do not require any special treatment.

The computing times within the tables are given in seconds, and are counting

- For the iterative method: the time needed for reaching convergence, or, in non-convergent experiments, until a maximum of 2000 iterations have been carried out.
- For the direct method: the time spent in defining the matrix plus the time devoted to solving the linear system.

When the seed for the iterative method is said to be computed with the (11, 10)-direct method, what is meant is that one uses the direct method with $I = 11$, $N = 10$ to solve the full problem, and then computes the seed, from the restriction of this solution to Q_0, by means of linear interpolation.

Lastly, the suffixes "IT" and "D" in the tables stand respectively, for "iterative" and "direct".

6.2. Test #1

This is a test taken from Subsection 3.2.1. One chooses

$$\alpha(\mu, z) = |\sin(12\mu z)|, \tag{6.1}$$

$$\sigma(\mu, z) = 1 + \sin(12\mu z)\cos(12\mu z), \tag{6.2}$$

$$\psi(\mu, z) = \ln(2 + \mu^2 + z^3), \tag{6.3}$$

and W, f and g are computed so that ψ is the exact solution.

6.2.1. Results obtained with the iterative method

Let us analyze some numerical results obtained with the iterative method. Results collected in Tables 6.2.1 and 6.2.2 show the influence that the choice of the seed has got in the number of iterations needed for convergence, while Table 6.2.3 shows that the number of iterations grows if no relaxation is done.

The meaning of "no convergence" in these three Tables 6.2.1, 6.2.2 and 6.2.3 is that convergence has not been reached yet, but the algorithm behaves in a convergent way; in other words, convergence is expected if more iterations are allowed. By way of example, we report in the last row of Table 6.2.2 the value of the residual after 2000 iterations.

$(I,\, N)$	$E_{\text{abs}}(Q)$-IT	iterations	time-IT (s)
$(11, 10)$	7.05×10^{-3}	42	0.17
$(33, 29)$	5.78×10^{-4}	136	1.83
$(101, 91)$	5.84×10^{-5}	405	19.01 (0.047 s/it)
$(321, 281)$	4.73×10^{-6}	1191	231.68 (0.195 s/it)
$(1001, 901)$	2.34×10^{-4}	No convergence in 2000 iterations	2212.43 (1.106 s/it)

Table 6.2.1: Numerical results for the test #1 by means of the iterative method with seed given by Equation *(5.8)*, $\omega = 2$ and $\varepsilon = 10^{-8}$.

(I, N)	$E_{\text{abs}}(Q)$-IT	iterations	time-IT (s)
$(11, 10)$	7.05×10^{-3}	1	0.0073
$(33, 29)$	5.78×10^{-4}	102	1.38
$(101, 91)$	5.90×10^{-5}	299	13.89 (0.046 s/it)
$(321, 281)$	6.78×10^{-6}	854	166.93 (0.195 s/it)
$(1001, 901)$	9.58×10^{-6}	No convergence in 2000 iterations (res.$=2.70 \times 10^{-8}$)	$\sim$ 2212 (1.106 s/it)

Table 6.2.2: Numerical results for the test #1 by means of the iterative method with seed given by the $(11, 10)$-direct method, $\omega = 2$ and $\varepsilon = 10^{-8}$.

$(I,\ N)$	$E_{\text{abs}}(Q)$-IT	iterations	time-IT(s)
$(11, 10)$	7.05×10^{-3}	1	0.0056
$(33, 29)$	5.78×10^{-4}	194	5.57
$(101, 91)$	5.94×10^{-5}	558	25.64 (0.046 s/it)
$(321, 281)$	7.87×10^{-6}	1568	303.19 (0.193 s/it)
$(1001, 901)$	2.06×10^{-4}	No convergence in 2000 iterations	$\sim$ 2212 (1.106 s/it)

Table 6.2.3: Numerical results for the test #1 by means of the iterative method with seed given by the $(11, 10)$-direct method , $\omega = 1$ (no relaxation) and $\varepsilon = 10^{-8}$.

In each of the Tables 6.2.1, 6.2.2 and 6.2.3, a constant value of the relaxation parameter ω has been used, but it is quite clear that results can be improved by choosing different values of ω for different meshes. With this spirit, Table 6.2.4 sets forth the number of iterations needed for convergence as a function of ω for three different meshes; this is the kind of analysis we have done for choosing the quasi-optimal relaxation parameters in Table 6.2.5, which is reporting the quasi-best results that one can expect to achieve with the iterative method.

	iterations with $(I, N) = (33, 29)$	iterations with $(I, N) = (101, 91)$	iterations with $(I, N) = (321, 281)$
$\omega = 1$	194	558	1568
$\omega = 1.5$	133	388	1100
$\omega = 2$	102	299	854
$\omega = 2.5$	82	244	701
$\omega = 3$	69	207	596
$\omega = 3.5$	59	180	520
$\omega = 4$	Divergence	159	461
$\omega = 5$	Divergence	129	377
$\omega = 6$	Divergence	Divergence	320
$\omega = 7$	Divergence	Divergence	278
$\omega = 8$	Divergence	Divergence	Divergence

Table 6.2.4: Number of iterations needed for convergence versus ω for test #1. The seed is given by the $(11, 10)$-direct method, and $\varepsilon = 10^{-8}$. "Divergence" is meaning blowing up of the numerical solution.

Notice that, although convergence has been reached, the result on the last row of Table 6.2.5 is not satisfactory, because the order 2 of the method is not being respected; in other words, a smaller value of ε must be used when solving with the (1001, 901)-mesh, which in turn means that more than 519 iterations are really needed to get the correct numerical solution.

$[(I,N),\,\omega]$	$E_{\text{abs}}(Q)$-IT	iterations	time-IT (s)
$[(11,10),\,2]$	7.05×10^{-3}	1	0.0057
$[(33,29),\,3.5]$	5.78×10^{-4}	59	0.76
$[(101,91),\,5]$	5.88×10^{-5}	129	6.20 (0.048 s/it)
$[(321,281),\,7]$	6.02×10^{-6}	278	53.82 (0.194 s/it)
$[(1001,\,901),\,11]$	1.18×10^{-6}	519	583.48 (1.124 s/it)

Table 6.2.5: Numerical results for the test #1 by means of the iterative method with seed given by the $(11,10)$-direct method, and $\varepsilon = 10^{-8}$. The value of ε is not small enough to keep accuracy in the last row (cf. Table 6.2.6).

In convergent cases, the iterative algorithm behaves like a globally convergent method. To give an instance, if we solve with $(I,N) = (101,91)$, $\omega = 5$, $\varepsilon = 10^{-8}$, and use the seed given by

$$\left(\psi_{i*}^{n}\right)^{[0]} = -10^{6} \text{ for } n \in \{2,\dots,N-1\}, \tag{6.4}$$

then convergence is achieved in 368 iterations (cf. Table 6.2.5). Observe that this seed is a long way away from the exact solution, the values of which on Q_0 are between $\ln(2)$ and $\ln(3)$.

6.2.2. Results obtained with the direct method

Table 6.2.6 is showing the results obtained with the direct method for several meshes. An expression like $t = {}_{t_1} + {}_{t_2}$ means that t_1 seconds were needed to define the matrix and t_2 seconds were needed to solve the linear system. By selecting the three rows (11, 10), (101, 91), (1001, 901) (or (33, 29), (321, 281), (3201, 2801)), the order 2 of the method can be seen by eye.

(I,N)	$E_{\text{abs}}(Q)$-D	time-D (s)
$(11,10)$	7.05×10^{-3}	$0.0030 = 0.0026 + 0.0004$
$(33,29)$	5.78×10^{-4}	$0.0098 = 0.0060 + 0.0038$
$(101,91)$	5.87×10^{-5}	$0.066 = 0.024 + 0.042$
$(321,281)$	5.72×10^{-6}	$0.80 = 0.23 + 0.57$
$(1001,901)$	5.85×10^{-7}	$11.26 = 1.93 + 9.33$
$(1501,1500)$	2.59×10^{-7}	$30.06 = 4.89 + 25.17$
$(2001,2000)$	1.46×10^{-7}	$68.14 = 8.76 + 59.38$
$(2501,2500)$	9.33×10^{-8}	$296.47 = 13.36 + 283.11$
$(3201,2801)$	5.72×10^{-8}	$1023.90 = 19.47 + 1004.43$

Table 6.2.6: Numerical results for the test #1 with the direct method.

When the computing times in Table 6.2.6 are compared with those in Table 6.2.5 the superiority of the direct method is apparent for meshes not finer than $(I, N) = (1001, 901)$, which by the way is fine enough to produce a very small error equal to 5.85×10^{-7}.

Since the linear systems which are solved when using the iterative algorithm are of order $\frac{I+1}{2}$ while the system solved when using the direct method is of order $I \times N$, it is clear that the superiority of the direct method cannot be maintained as the mesh is refined. The point is that the range of meshes where the direct method is faster contains the meshes that are typically used, and that this fact has been persistently ignored.

Let us comment now on the meshes finer than $(I,\ N) = (1001,\ 901)$. We have checked that one iteration takes 8.55 seconds if $(I,\ N) = (3201,\ 2801)$, and hence even for this fine mesh the iterative algorithm should converge in no more than $\frac{1023.9}{8.55} \approx 119$ iterations to be faster than the direct method; this is very optimistic: indeed, the residual (with seed given by the (11, 10)-direct method and $\omega = 12$) after 119 iterations is equal to 1.32×10^{-5}, far away from the tolerance, which must be less than 10^{-8} to take profit of this fine mesh. This reasoning proves that the direct method is the fastest at least for all meshes in Table 6.2.6.

6.3. Test #2

Again, this is a test taken from Subsection 3.2.2. One chooses

$$\alpha(\mu, z) = 0.02,\ \sigma(\mu, z) = 0.01,\ W(\mu, z) = 0, \tag{6.5}$$

$$f(\mu) = 1,\ g(\mu) = 2. \tag{6.6}$$

In this case, the exact solution is not explicitly known, is continuous and, despite the C^∞ regularity of the data, has got singularities in some of its partial derivatives of first order at $(\mu, z) \in \{(0,\ 0),\ (0,\ 1)\}$. This lack of regularity prevents the scheme from having order 2 when applied to this problem. Plots and numerical results can be found in Subsection 3.2.2.

Even when, regardless the data functions, the patterns of the matrices are the same for the same mesh, one can observe some differences in the computing times between this test and the previous one. To explain this phenomenon, simply notice that LU factorization does pivoting based not only on the pattern but also on the values as well. So changing the values can easily change the pivot order, which in turn can affect the time (and also the amount of memory) taken.[j]

[j] Prof. Tim Davis personal communication.

6.3.1. Results obtained with the iterative method

Tables 6.3.1 and 6.3.2 are analogous to Tables 6.2.4 and 6.2.5, and must be interpreted alike.

	iterations with $(I, N) = (33, 29)$	iterations with $(I, N) = (101, 91)$	iterations with $(I, N) = (321, 281)$
$\omega = 1$	42	117	322
$\omega = 1.5$	28	80	223
$\omega = 1.7$	30	71	199
$\omega = 1.8$	46	68	189
$\omega = 1.9$	89	66	180
$\omega = 2$	1497	220	171
$\omega = 2.1$	Divergence	Divergence	164
$\omega = 2.2$	Divergence	Divergence	166
$\omega = 2.3$	Divergence	Divergence	Divergence

Table 6.3.1: Number of iterations needed for convergence versus ω for test #2. The seed is given by the $(11, 10)$-direct method, and $\varepsilon = 10^{-8}$. "Divergence" is meaning blowing up of the numerical solution.

$[(I, N), \omega]$	iterations	time-IT (s)
$[(11, 10), 2]$	1	0.0057
$[(33, 29), 1.5]$	28	0.31
$[(101, 91), 1.9]$	66	2.41 (0.037 s/it)
$[(321, 281), 2.1]$	164	20.31 (0.124 s/it)
$[(1001, 901), 2.7]$	367	190.19 (0.518 s/it)

Table 6.3.2: Numerical results for the test #2 by means of the iterative method with the seed given by the $(11, 10)$-direct method, and $\varepsilon = 10^{-8}$.

6.3.2. Results obtained with the direct method

Once more, results in Tables 6.3.2 and 6.3.3 show that the direct method outperforms the iterative one. For this test, one iteration with $(I, N) = (3201, 2801)$ takes about 2.90 seconds, and arguments already used to prove that the direct method is faster for this mesh, and hence for all meshes in Table 6.3.3.

(I, N)	time-D (s)
$(11, 10)$	$0.0070 = 0.0032 + 0.0038$
$(33, 29)$	$0.014 = 0.007 + 0.007$
$(101, 91)$	$0.062 = 0.015 + 0.047$
$(321, 281)$	$0.77 = 0.13 + 0.64$
$(1001, 901)$	$14.67 = 1.27 + 13.40$
$(1501, 1500)$	$45.34 = 3.19 + 42.15$
$(2001, 2000)$	$101.15 = 6.80 + 94.35$
$(2501, 2500)$	$381.52 = 9.62 + 371.90$
$(3201, 2801)$	$1109.50 = 12.85 + 1096.65$

Table 6.3.3: Numerical results for the test #2 with the direct method.

6.4. Some additional comments

It has been already explained in Section 6.3 that asymptotically, which in this case is meaning "as the mesh is being more and more refined", the iterative algorithm will always win from a certain fineness onwards. However, from a practical perspective, the question one must answer is whether the meshes which are going to be used are among those which make the direct method be the fastest or not. In relation to this standpoint, our experiments indicate that the direct method is the one that must be used for a pretty wide range of useful meshes.

We detail in the following points several arguments that play in favour of the direct method in the framework we are:

1. The number of iterations needed for convergence depends strongly on the value of the relaxation parameter ω, but the optimum value of this parameter is not known in advance and, moreover, is different for different problems.

2. When one refines the mesh, one should obtain a better, more accurate solution. In order to compute the solution with the expected accuracy, the tolerance parameter ε might have to be reduced, which would increase the number of iterations. On the other hand, the direct method does not suffer from this inconvenience.

3. Loops cannot be avoided when coding the iterative algorithm: one external loop in iterations, one internal loop for solving on Q_+, and one internal loop for solving on Q_-. On the contrary, loops are not needed to code the direct method.

Chapter 7

Azimuthal variable dependent problem

Contents

7.1. Introduction **80**

7.2. Hypotheses **81**

7.3. Fourier technique **84**

7.4. Scheme for the θ-independent problem *(7.35)–(7.37)* **86**

7.4.1. Description of the core scheme 88

7.4.2. Numerical experiments with the core scheme 90

7.5. Numerical results for the full problem **92**

7.5.1. Test #1 93

7.5.2. Test #2 94

7.1. Introduction

This Chapter is devoted to solving the steady monoenergenitic Fokker-Planck equation in the 1D slab (see Figure 1.1) when the angular flux ψ also depends on the azimuthal angle θ. Fourier techniques are employed to split the problem into a collection of θ-independent problems whose absorption coefficient becomes singular. The presence of these singularities obliges to modify the odd scheme introduced in Section 4.1 in order to use it also in this case. Numerical experiments conducted and discussed show second order of convergence.

Let $Q_{\mu,z,\theta} \subset \mathbb{R}^3$ be the open set $(-1,1) \times (Z_{\text{ini}}, Z_{\text{fin}}) \times (0, 2\pi)$ and let $\overline{Q}_{\mu,z,\theta}$ be the closure, that is, $[-1,1] \times [Z_{\text{ini}}, Z_{\text{fin}}] \times [0, 2\pi]$. The notation (μ, z, θ) will stand for a generic element of any of these sets, what gives to $Q_{\mu,z}$ and other choices of subindexes in $\{\mu, z, \theta\}$ must be understood analgously.

We seek a function

$$\psi : (\mu, z, \theta) \in \overline{Q}_{\mu,z,\theta} \rightarrow \psi(\mu, z, \theta) \in \mathbb{R}$$

which is the solution of the partial differential equation (PDE)

$$\mu \frac{\partial \psi}{\partial z} + \alpha\psi - \sigma \left\{ \frac{\partial}{\partial \mu} \left[(1-\mu^2) \frac{\partial \psi}{\partial \mu} \right] + \frac{1}{1-\mu^2} \frac{\partial^2 \psi}{\partial \theta^2} \right\} = W \tag{7.1}$$

in $Q_{\mu,z,\theta}$, accompanied by the following conditions:

$$\psi_{|\{\mu\in(0,1],\, z=Z_{\text{ini}}\}} = f, \text{ with } f = f(\mu,\theta) \text{ given}, \tag{7.2}$$

$$\psi_{|\{\mu\in[-1,0),\, z=Z_{\text{fin}}\}} = g, \text{ with } g = g(\mu,\theta) \text{ given}, \tag{7.3}$$

$$\psi_{|\{\theta=0\}} = \psi_{|\{\theta=2\pi\}}, \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=0\}} = \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=2\pi\}}. \tag{7.4}$$

Function ψ is representing the angular flux density of charged particles, for example electrons. Variable z stands for 1D space, μ for the cosine of the polar angle and θ for the azimuthal angle.[(k)] In the their maximal generality, functions α, σ and W could depend on (μ, z, θ), but the technique used in this Chapter does not allow functions α and σ to depend on θ; in other words, α and σ depend only on (μ, z). Function W, on the other hand, can depend on (μ, z, θ).

We refer to Section 1.2 for the physical domain on which the problem *(7.1)–(7.4)* is considered. Simply consider the same explanation with the caveat that now the continuous scattering operator is

$$\frac{\partial}{\partial \mu} \left[(1-\mu^2) \frac{\partial \cdot}{\partial \mu} \right] + \frac{1}{1-\mu^2} \frac{\partial^2 \cdot}{\partial \theta^2}. \tag{7.5}$$

It has the form $\frac{\partial}{\partial \mu} \left[(1-\mu^2) \frac{\partial \cdot}{\partial \mu} \right]$ for the θ-independent problem studied in Chapters 3-6.

We remark that the scattering operator *(7.5)* is the Laplacian over the sphere, also, called the Laplace-Beltrami operator, which connects this problem with the interesting field of surface PDEs. This is, nevertheless, an approach that we have not pursued in this work.

[(k)] One must keep in mind the standard spherical coordinate system at a generic point $(x_1, x_2, z) \in \mathbb{R}^3$. Since the 1D slab assumes independency of x_1 and x_2, it is sufficient to work, from the spatial standpoint, with the z-coordinate.

7.2. Hypotheses

Let

$$W_0(\mu, z) = \frac{1}{2\pi}\int_0^{2\pi} W(\mu, z, \theta)d\theta \tag{7.6}$$

be the zeroth Fourier coefficient of W with respect to θ. It will be assumed that

$$f \in \mathrm{C}^\infty\big([0,1]\times\overline{Q}_\theta\big), \tag{7.7}$$
$$g \in \mathrm{C}^\infty\big([-1,0]\times\overline{Q}_\theta\big), \tag{7.8}$$
$$\alpha,\ \sigma \in \mathrm{C}^\infty\big(\overline{Q}_{\mu,z}\big), \tag{7.9}$$
$$\alpha \geq 0,\ \ \sigma > 0, \tag{7.10}$$
$$W_0 \in \mathrm{C}^\infty\big(\overline{Q}_{\mu,z}\big),\ \ (1-\mu^2)W \in \mathrm{C}^\infty\big(\overline{Q}_{\mu,z,\theta}\big), \tag{7.11}$$
$$f,\ g \text{ and } W \text{ are periodic in the sense of Equation (7.4).} \tag{7.12}$$

We notice that under the hypotheses above

- $f(\mu,\cdot)$, for every $\mu \in [0,1]$,
- $g(\mu,\cdot)$, for every $\mu \in [-1,0]$, and
- $W(\mu, z, \cdot)$, for every $(\mu, z) \in Q_\mu \times \overline{Q}_z$,

are all representable by their Fourier series with respect to the variable θ (for a justification, see for instance [31]), a fact which will be of importance in this Chapter.

Sometimes, W will not be defined at $\mu = \pm 1$. In this case, hypotheses (7.11) means that functions $(1-\mu^2)W$ and W_0 admit of C^∞ extensions to $\overline{Q}_{\mu,z,\theta}$ and $\overline{Q}_{\mu,z}$, respectively, and the same notations will be used for the extended functions.

To give an instance, the function

$$W(\mu, z, \theta) = \mu z + \frac{\sin(\theta)}{1-\mu^2}$$

satisfies (7.11) but the following one

$$W(\mu, z, \theta) = \mu z + \frac{\sin(\theta/2)}{1-\mu^2}$$

does not. Consider, $W(\mu, z, \theta) = \mu z + \frac{\sin(\theta/2)}{1-\mu^2}$. Since

$$\int_0^{2\pi} \sin(\theta/2)d\theta = -2\cos(\theta/2)\big|_0^{2\pi} = 4, \tag{7.13}$$

one has got $W_0(\mu, z) = \frac{1}{2\pi}(2\pi\mu z + \frac{4}{1-\mu^2}) \in \mathrm{C}^\infty(\overline{Q}_\mu \times \overline{Q}_z)\backslash\mathrm{C}^\infty(\overline{Q}_{\mu,z})$.

It is clear that condition (7.11) holds for any $W \in \mathrm{C}^\infty(\overline{Q}_{\mu,z,\theta})$. The following lemma gives a necessary and sufficient condition for (7.11) to hold; its proof is a simple exercise.

The notation $(\mathcal{H}_W)_0$ stands, as in Equation *(7.6)*, for the zeroth Fourier coefficient of $\mathcal{H}_W$ with respect to θ.

Lemma 7.2.1. *W satisfies (7.11) if, and only if, it admits the representation*

$$W(\mu,z,\theta)=\widehat{W}(\mu,z)+\frac{\mathcal{H}_W(\mu,z,\theta)}{1-\mu^2},\ (\mu,z,\theta)\in\left(Q_\mu\times\overline{Q}_z\times\overline{Q}_\theta\right), \tag{7.14}$$

being, on the one hand, $\widehat{W}$ a function in $\mathrm{C}^\infty\left(\overline{Q}_{\mu,z}\right)$ and on the other hand, $\mathcal{H}_W$ a function in $\mathrm{C}^\infty\left(\overline{Q}_{\mu,z,\theta}\right)$ such that $\left(\mathcal{H}_W\right)_0\equiv 0$ in $\overline{Q}_{\mu,z}$.

Necessarily, $\widehat{W}=W_0$ and $\mathcal{H}_W=(1-\mu^2)(W-W_0)$, and so the representation (7.14) is unique.

In this Section we will perform various mathematical operations involving ψ. All of them are valid if

$$\psi\in\mathrm{C}(\overline{Q}_{\mu,z,\theta}), \tag{7.15}$$

$$\frac{\partial\psi}{\partial z},\ \frac{\partial^2\psi}{\partial\mu^2}\in\mathrm{C}(Q_z\times\overline{Q}_{\mu,\theta}), \tag{7.16}$$

$$\frac{\partial^2\psi}{\partial\theta^2}\in\mathrm{C}(\overline{Q}_{\mu,z,\theta}). \tag{7.17}$$

The type of regularity expressed by *(7.15)–(7.17)* is in accordance with the numerical results, which support the idea that, under assumptions *(7.7)–(7.12)*, ψ is continuous on $\overline{Q}_{\mu,z,\theta}$ and smooth on $Q_z\times\overline{Q}_{\mu,\theta}$. The spatial domain *(7.16)* is open in order to allow the occurrence of the following phenomenon, which has been experimentally observed: even for constant data functions, $\frac{\partial\psi}{\partial z}$ can be infinity and $\frac{\partial\psi}{\partial\mu}$ can fail to exist at (μ,z,θ) when $(\mu,z)\in\{(0,Z_{\mathrm{ini}}),(0,Z_{\mathrm{fin}})\}$.

Remark 7.2.1. *As it happens for other problems, the existence of solutions satisfying regularity properties can depend on certain compability conditions between the data. This will become clear later.*

Class C^p regularity of ψ with respect to θ with $p>2$, if it happens, makes the method described in this Chapter be more efficient. This is because ψ is approximated by a truncation of its Fourier series, whose speed of convergence grows with the regularity.

The Physics of this problem demands that ψ be non-negative and since the values $\mu\pm 1$ represent the same points on the sphere (North and South poles) for every θ, it also demands that $\psi(\pm 1,\cdot,\cdot)$ do not depend on θ. In order that ψ be non-negative, f and g must be non-negative as well, and W, which has not got sign restrictions, can be negative only up to a point that still render $\psi\geq 0$. The following theorem provides us with a necessary and sufficient condition for $\psi(\pm 1,\cdot,\cdot)$ to be independent of θ.

Theorem 7.2.1. *Let ψ be the solution of the problem (7.1)–(7.4). Under the hypotheses (7.7)–(7.12) and assuming that conditions (7.15)–(7.17) are satisfied, the two functions $\psi(\pm 1,\cdot,\cdot)$ to be independent of θ if, and only if, the following two assertions hold simultaneously:*

$$\textit{Functions } f(1,\cdot),\ \textit{and } g(-1,\cdot)\ \textit{do not depend on } \theta\ \textit{and} \tag{7.18}$$

$$\textit{Functions } \mathcal{H}_W(\pm 1,\cdot,\cdot) \textit{ are identicaly } 0. \tag{7.19}$$

Proof. Firstly let us multiply the PDE (7.1) by $1-\mu^2$, and then take limits as $\mu \to \pm 1$ and employ continuity to get, using Equation (7.14),

$$\sigma(\pm 1, z)\frac{\partial^2 \psi}{\partial \theta^2}(\pm 1, z, \theta) = \mathcal{H}_W(\pm 1, z, \theta),\ (z,\theta) \in \overline{Q}_{z,\theta}. \tag{7.20}$$

Now the proof follows easily by taking into account conditions (7.2)–(7.4). □

Examples:

- $f(\mu,\theta) = 1 + (1-\mu)\sin\theta$.
- $g(\mu,\theta) = 2 + (1+\mu)\cos\theta$.
- If $W(\mu, z, \theta) = 3 + \mathrm{e}^{\mu z} + z^3 \sin\theta\cos\theta$, then

  $$W(\mu, z, \theta) = 3 + \mathrm{e}^{\mu z} + \frac{z^3(1-\mu^2)\sin\theta\cos\theta}{1-\mu^2}$$

  i. e., $\widehat{W}(\mu, z) = 3 + 3 + \mathrm{e}^{\mu z}$ and $\mathcal{H}_W(\mu, z, \theta) = z^3(1-\mu^2)\sin\theta\cos\theta$. These functions satisfy all the conditions of Lemma 7.2.1.
- But, if we get $W(\mu, z, \theta) = 3 + \mathrm{e}^{\mu z} + \frac{z^3 \sin\theta\cos\theta}{\sqrt{1-\mu^2}}$, then

  $$W(\mu, z, \theta) = 3 + \mathrm{e}^{\mu z} + \frac{z^3\sqrt{1-\mu^2}\sin\theta\cos\theta}{1-\mu^2},$$

  i. e., $\widehat{W}(\mu, z) = 3 + \mathrm{e}^{\mu z}$ and $\mathcal{H}_W(\mu, z, \theta) = z^3\sqrt{1-\mu^2}\sin\theta\cos\theta$. The function $\mathcal{H}_W(\mu, z, \theta)$ does not belong to $\mathrm{C}^\infty\left(\overline{Q}_{\mu,z,\theta}\right)$.

Remark 7.2.2. *It is clear that condition (7.18) is exactly meaning that functions $f(1,\cdot)$ and $g(-1,\cdot)$ are constant.*

Remark 7.2.3. *(Compatibility conditions). Equations (7.20) can be further exploited to assert that, under the hypotheses (7.7)–(7.12), a necessary condition for the problem (7.1)–(7.4) to have got a solution satisfying conditions (7.15)–(7.17) is that the following two equalities be satisfied for $\forall\theta \in \overline{Q}_\theta$:*

$$f(1,\theta) = f(1,0) + \frac{1}{\sigma(1, Z_{\text{ini}})}\left\{\frac{\theta}{2\pi}\int_0^{2\pi}\left(\int_0^s \mathcal{H}_W(1, Z_{\text{ini}}, t)dt\right)ds - \int_0^\theta\left(\int_0^s \mathcal{H}_W(1, Z_{\text{ini}}, t)dt\right)ds\right\}, \tag{7.21}$$

$$g(-1,\theta) = g(-1,0) + \frac{1}{\sigma(-1, Z_{\text{fin}})}\left\{\frac{\theta}{2\pi}\int_0^{2\pi}\left(\int_0^s \mathcal{H}_W(-1, Z_{\text{fin}}, t)dt\right)ds - \int_0^\theta\left(\int_0^s \mathcal{H}_W(-1, Z_{\text{fin}}, t)dt\right)ds\right\}, \tag{7.22}$$

Notice that both of them hold if conditions (7.18) and (7.19) are fulfilled.
In order to prove (7.21) and (7.22), firstly integrate Equation (7.20) between 0 and θ to obtain

$$\frac{\partial\psi}{\partial\theta}(\pm 1, z, \theta) = \frac{\partial\psi}{\partial\theta}(\pm 1, z, 0) - \frac{1}{\sigma(\pm 1, z)}\int_0^\theta \mathcal{H}_W(\pm 1, z, s)ds. \tag{7.23}$$

We can remark here that $\frac{\partial\psi}{\partial\theta}(\pm1,z,0)=\frac{\partial\psi}{\partial\theta}(\pm1,z,2\pi)$ *is guaranteed by the property* $(\mathcal{H}_W)_0\equiv 0$ *(see Lemma 7.2.1).*
Secondly, integrate Equation (7.23) between 0 *and* θ *and take into account that* $\psi_{|\{\theta=0\}}=\psi_{|\{\theta=2\pi\}}$ *to prove that*

$$\psi(\pm1,z,\theta)=\psi(\pm1,z,0)+\frac{1}{\sigma(\pm1,z)}\left\{\frac{\theta}{2\pi}\int_0^{2\pi}\left(\int_0^s\mathcal{H}_W(\pm1,z,t)dt\right)ds-\int_0^{\theta}\left(\int_0^s\mathcal{H}_W(\pm1,z,t)dt\right)ds\right\}. \tag{7.24}$$

If we have not added to the hypotheses (7.7)–(7.12) either conditions guaranteeing non-negativity of ψ or the conditions (7.18) and (7.19) listed in the Theorem 7.2.1, it is because, from a mathematical viewpoint, one might want to solve the problem also in a non-physical framework.

Some of the hypotheses (7.7)–(7.12) can be weakened, and the reader will find within the numerical results and examples where some of them are failing while at the same time the numerical method is properly working.

7.3. Fourier technique

Periodic conditions (7.4) invite to use Fourier techniques to reduce the problem (7.1)–(7.4) to a collection of θ-independent problems (see also [21]). Let us consider the Fourier series of ψ with respect to θ, that is to say,

$$\psi(\mu,z,\theta)=\sum_{k=-\infty}^{\infty}\psi_k(\mu,z)\,\mathrm{e}^{ik\theta}, \tag{7.25}$$

$(\mu,z,\theta)\in\overline{Q}_{\mu,z,\theta}$, where the coefficients ψ_k are given by the well-known formula

$$\psi_k(\mu,z)=\frac{1}{2\pi}\int_0^{2\pi}\psi(\mu,z,\theta)\,\mathrm{e}^{-ik\theta}d\theta,\quad(\mu,z)\in\overline{Q}_{\mu,z}. \tag{7.26}$$

From Equation (7.26) it is clear that $\psi_{-k}=\overline{\psi_k}$ for every $k\in\mathbb{Z}$, because the function ψ is real. As a result, Equation (7.25) can be written as

$$\psi(\mu,z,\theta)=\psi_0(\mu,z)+2\sum_{k=1}^{\infty}\Big\{\mathrm{Re}[\psi_k(\mu,z)]\cos(k\theta)-\mathrm{Im}[\psi_k(\mu,z)]\sin(k\theta)\Big\}. \tag{7.27}$$

For numerical purposes, Equation (7.27) is more suitable than Equation (7.25), because it prevents spurious complex non-real values from appearing. Naturally, one must understand that the series will be truncated at a certain level. In case that the series converges rapidly, one can get ψ with high accuracy from the knowledge of only a few coefficients ψ_k. Once the commanding harmonics have been reached, rapid convergence occurs for instance when ψ is, for each fixed pair (μ,z), the restriction to $[0,2\pi]$ of a (2π)-periodic function of class $C^p(\mathbb{R})$; the larger p, the faster rate of convergence (see for instance [31]).

The point is that each ψ_k is the solution to a Fokker-Planck equation much cheaper than Equation *(7.1)* because it is free of θ-dependence. In order to obtain this equation, the following notation is going to be used in this Section: G_k will denote the k^{th} Fourier coefficient of a function G with respect to the variable θ.

Now the first thing to do, according to Equation *(7.26)*, is to multiply each term in Equation *(7.1)* by $\mathrm{e}^{-ik\theta}$, then integrate from $\theta = 0$ to $\theta = 2\pi$, and finally divide by 2π. The results are the following:

$$\frac{1}{2\pi}\int_0^{2\pi} \mu\frac{\partial\psi}{\partial z}\,\mathrm{e}^{-ik\theta}d\theta = \mu\frac{\partial\psi_k}{\partial z}, \tag{7.28}$$

$$\frac{1}{2\pi}\int_0^{2\pi} \alpha\psi\,\mathrm{e}^{-ik\theta}d\theta = \alpha\psi_k, \tag{7.29}$$

$$\frac{1}{2\pi}\int_0^{2\pi} \sigma\frac{\partial}{\partial\mu}\left[(1-\mu^2)\frac{\partial\psi}{\partial\mu}\right]\mathrm{e}^{-ik\theta}d\theta = \sigma\frac{\partial}{\partial\mu}\left[(1-\mu^2)\frac{\partial\psi_k}{\partial\mu}\right], \tag{7.30}$$

$$\frac{1}{2\pi}\int_0^{2\pi} \frac{\sigma}{1-\mu^2}\frac{\partial^2\psi}{\partial\theta^2}\,\mathrm{e}^{-ik\theta}d\theta = -\frac{\sigma k^2}{1-\mu^2}\psi_k, \tag{7.31}$$

$$\frac{1}{2\pi}\int_0^{2\pi} W\,\mathrm{e}^{-ik\theta}d\theta = W_k. \tag{7.32}$$

Under the hypotheses *(7.7)–(7.12)* and assuming that conditions *(7.15)–(7.17)* are satisfied, we can state that Equations *(7.28)–(7.32)* hold for all $(\mu, z) \in Q_{\mu,z}$. The theorem we are thinking of for differentiating under the integral sign is [7, (8.11.2), p.177].

Looking at Equations *(7.29)*, *(7.30)* and *(7.31)* it is now understood why, as anticipated, neither α nor σ can depend on θ.

Equation *(7.31)*, the only one which might demand further explanation, is the result of integrating twice by parts and taking into account the periodic conditions *(7.4)*:

$$\begin{aligned}\int_0^{2\pi} \frac{\partial^2\psi}{\partial\theta^2}\,\mathrm{e}^{-ik\theta}d\theta = \left(\frac{\partial\psi}{\partial\theta}\right)_{|\{\theta=2\pi\}} - \left(\frac{\partial\psi}{\partial\theta}\right)_{|\{\theta=0\}} + \\ ik(\psi_{|\{\theta=2\pi\}} - \psi_{|\{\theta=0\}}) - k^2\int_0^{2\pi}\psi\,\mathrm{e}^{-ik\theta}d\theta = \\ -k^2\int_0^{2\pi}\psi\,\mathrm{e}^{-ik\theta}d\theta.\end{aligned} \tag{7.33}$$

Now, adding Equations *(7.28)–(7.29)* on the one hand, Equations *(7.30)–(7.31)* on the other hand, using Equation *(7.1)*, and defining

$$\mathtt{A}(k) := \alpha + \frac{k^2\sigma}{1-\mu^2}, \tag{7.34}$$

one arrives at the problem of solving for

$$\psi_k : (\mu, z) \in \overline{Q}_{\mu,z} \to \psi_k(\mu, z) \in \mathbb{C}$$

the PDE

$$\mu\frac{\partial\psi_k}{\partial z} + \mathtt{A}(k)\psi_k = \sigma\frac{\partial}{\partial\mu}\left[(1-\mu^2)\frac{\partial\psi_k}{\partial\mu}\right] + W_k \tag{7.35}$$

in $Q_{\mu,z}$, accompanied by the conditions

$$\psi_k(\mu, Z_{\text{ini}}) = f_k(\mu) \text{ for } \mu \in (0,1], \tag{7.36}$$
$$\psi_k(\mu, Z_{\text{fin}}) = g_k(\mu) \text{ for } \mu \in [-1,0). \tag{7.37}$$

Remark 7.3.1. *When $k = 0$, one faces a problem like the one studied in [25]. However, the "absorption" coefficient $\mathbf{A}(k)$ is singular at $\mu = \pm 1$ if $k > 0$, what distinguish these cases from the former one.*

Remark 7.3.2. *According to Equation (7.14), one has, when $k > 0$,*

$$W_k(\mu, z) = \frac{(\mathcal{H}_W)_k(\mu, z)}{1-\mu^2} \quad \forall (\mu, z) \in Q_\mu \times \overline{Q}_z. \tag{7.38}$$

Notice from Equation (7.27) that $\psi(\pm 1, \cdot, \cdot)$ is independent of θ if, and only if,

$$\psi_k(\pm 1, \cdot, \cdot) \text{ is identically 0 for all } \quad k > 0. \tag{7.39}$$

The values $\psi_0(\pm 1, \cdot)$ cannot be derived *a priori* and are obtained as a by-product of the resolution process. The following theorem gives a direct way for computing $\psi_k(\pm 1, \cdot)$ when $k > 0$.

Theorem 7.3.1. *Under the hypotheses (7.7)–(7.12) and assuming that conditions (7.15)–(7.17) are satisfied, one has, when $k > 0$,*

$$\psi_k(\pm 1, z) = \frac{(\mathcal{H}_W)_k(\pm 1, z)}{k^2 \sigma(\pm 1, z)} \quad \forall z \in \overline{Q}_z. \tag{7.40}$$

Proof. Notice firstly that ψ_k inherits the regularity of ψ. Now fix $k > 0$, multiply the PDE in Equation (7.35) by $(1-\mu^2)$, take limits as $\mu \to \pm 1$ and employ continuity in order to conclude, using (7.14)–(7.15), that (7.40) holds. □

Remark 7.3.3. *(Compatibility conditions). Theorem 7.3.1 implies that, under the hypotheses (7.7)–(7.12), it is necessary that*

$$k^2 \sigma(1, Z_{\text{ini}}) f_k(1) = (\mathcal{H}_W)_k(1, Z_{\text{ini}}), \tag{7.41}$$
$$k^2 \sigma(-1, Z_{\text{fin}}) g_k(-1) = (\mathcal{H}_W)_k(-1, Z_{\text{fin}}) \tag{7.42}$$

for all $k > 0$ so that the problem (7.1)–(7.4) have got a solution satisfying conditions (7.15)–(7.17).

7.4. Scheme for the θ-independent problem (7.35)–(7.37)

In this Section, for the sake of brevity, we will not mark out k-dependencies, so that there be no risk of confusion in using later on the same letter k to indicate the distance between z-nodes and moreover, i will be a natural subindex for μ-nodes, rather than the imaginary unit.

A numerical scheme for solving a slightly more general problem than *(7.34)–(7.37)* will be described. Consider a function $\mathcal{H}(\mu, z) \geq 0$, continuous on $[-1, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}]$ and assume that $\mathtt{A} : (-1, 1) \times [Z_{\text{ini}}, Z_{\text{fin}}] \to [0, \infty)$ can be written as

$$\mathtt{A}(\mu, z) = \frac{\mathcal{H}(\mu, z)}{1 - \mu^2} \tag{7.43}$$

Only the following two cases are in our mind, because they are enough to cover all possibilities represented by Equation *(7.34)*: we shall call the *regular case* the one we have when $\mathtt{A}$ can be extended with continuity to $[-1, 1]\times[Z_{\text{ini}}, Z_{\text{fin}}]$ and we shall call the *singular case* the one we have when, for every $z, \mathcal{H}(-1, z)$ and $\mathcal{H}(1, z)$ are positive ("positive" is meaning "strictly positive"). In the regular case, the same notation $\mathtt{A}$ will be used for the extended function.

By way of illustration, the case $k = 0$ in Equation *(7.34)* is regular, while all cases with $k > 0$ are singular. Hypotheses *(7.9)* and *(7.10)* are playing a role in these assertions.

According to the discussion in the previous Section, it suffices, in order to compute the Fourier coefficients of the angular flux, to have a numerical scheme for the *core problem* consisting of the PDE

$$\mu \frac{\partial \psi}{\partial z} + \mathtt{A}\psi = \sigma \frac{\partial}{\partial \mu}\left[(1 - \mu^2)\frac{\partial \psi}{\partial \mu}\right] + W, \tag{7.44}$$

$$(\mu, z) \in \overline{Q}_{\mu,z} = [-1, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}]$$

and the boundary conditions

$$\psi(\mu, Z_{\text{ini}}) = f(\mu) \text{ for } \mu \in (0, 1], \tag{7.45}$$

$$\psi(\mu, Z_{\text{fin}}) = g(\mu) \text{ for } \mu \in [-1, 0). \tag{7.46}$$

The idea for solving the core problem is to use the odd scheme described in Section 4.1, which as it stands is usable only in the regular case, including a correction that capacitates it to solve the singular case as well. It has interest to consider the possibility that not only $\mathtt{A}$ but also W can be singular; the reason is that sometimes W inherits the singularities of $\mathtt{A}$ when one starts from a known exact solution ψ and adjusts W at the end so that Equation *(7.44)* be satisfied; also, some tests where W is singular by its own have been successfully conducted and will be shown later. There is however a difference between the singularities of $\mathtt{A}$ and those of W, because only the first ones unavoidably arise, as it is stated by Equation *(7.34)*, in the course of the resolution process even when all data functions in the original problem *(7.1)–(7.4)* be regular.

Let us mention to finish the introduction of this Section that sometimes the value of the solution at $|\mu| = 1$ can be inferred from the data functions $\mathtt{A}$ and W when the singular case is being considered. In such a case, the values of $f(1)$ and $g(-1)$ have to satisfy compatibility conditions so that a continuous solution ψ exist. Indeed, it can be proved by similar arguments to those used in the introduction that, under certain reasonable assumptions, one has $\psi(\pm 1, \cdot) \equiv 0$, which obliges to $f(1) = g(-1) = 0$, if

$$\lim_{\mu \to \pm 1} (1 - \mu^2) W(\mu, \cdot) = 0, \text{ pointwise} \tag{7.47}$$

or, more generally, that one has $\psi(\pm 1, z) = \frac{\zeta_\pm(z)}{\mathcal{H}(\pm 1, z)}$ for all $z \in [Z_{\rm ini}, Z_{\rm fin}]$, which obliges to $f(1) = \frac{\zeta_+(Z_{\rm ini})}{\mathcal{H}(1, Z_{\rm ini})}$ and $g(-1) = \frac{\zeta_-(Z_{\rm fin})}{\mathcal{H}(-1, Z_{\rm fin})}$, if

$$\lim_{\mu \to \pm 1} (1 - \mu^2) W(\mu, \cdot) = \zeta_\pm, \text{ pointwise,} \tag{7.48}$$

with $\zeta_\pm$ two functions in $\mathrm{C}\big([Z_{\rm ini}, Z_{\rm fin}]\big)$.

According to *(7.39)* and comments in the Introduction, it is condition *(7.47)* the one that respects the geometrical meaning of the variable θ.

The expression "core scheme" in the next Subsection will make reference to the scheme for solving the core problem *(7.44)–(7.46)*.

7.4.1. Description of the core scheme

As in Section 3.1, let us consider the uniform meshes

$$\mu_i = -1 + (i-1)h \text{ for } i \in \{1, \dots, I\}, \text{ with } h = \Delta\mu = \frac{2}{I-1} \tag{7.49}$$

and

$$z_n = Z_{\rm ini} + (n-1)k \text{ for } k \in \{1, \dots, N\}, \text{ with } k = \Delta z = \frac{Z_{\rm fin} - Z_{\rm ini}}{N-1}. \tag{7.50}$$

Hence,

$$\mu_1 = -1 < \mu_2 < \cdots < \mu_{I-1} < \mu_I = 1 \tag{7.51}$$

and

$$z_1 = Z_{\rm ini} < z_2 < \cdots < z_{N-1} < z_N = Z_{\rm fin}. \tag{7.52}$$

Actually, two schemes, both order 2, are described in Subsection 3.1.2 and Section 4.1: the even and the odd schemes. Numerical results obtained in Section 3.2 using these schemes clearly show that the second one is better behaved, it is the odd scheme the one to be used. This means that I is odd and, consequently, $\mu_{i^*} = 0$ if $i^* = \frac{I+1}{2}$.

The following notations will be employed: $\overline{D}_i = D(\mu_i)$ and $\overline{D}_{i\pm\frac{1}{2}} = D(\mu_i \pm \frac{h}{2})$, being $D(\mu) = 1 - \mu^2$; $\overline{\mathtt{A}}_i^n = \mathtt{A}(\mu_i, z_n)$, $\overline{\sigma}_i^n = \sigma(\mu_i, z_n)$, $\overline{W}_i^n = W(\mu_i, z_n)$; $\overline{f}_i = f(\mu_i)$, $\overline{g}_i = g(\mu_i)$; $\psi_i^n \approx \psi(\mu_i, z_n)$.

Morever,

$$\widetilde{\mathtt{A}}_1^n = \widetilde{\mathtt{A}}_1^n(h) = \begin{cases} \overline{\mathtt{A}}_1^n \text{ in the regular case,} \\ \mathtt{A}(\mu_1 + h^4, z_n) \text{ in singular case,} \end{cases} \tag{7.53}$$

$$\widetilde{\mathtt{A}}_I^n = \widetilde{\mathtt{A}}_I^n(h) = \begin{cases} \overline{\mathtt{A}}_I^n \text{ in the regular case,} \\ \mathtt{A}(\mu_I - h^4, z_n) \text{ in singular case,} \end{cases} \tag{7.54}$$

and

$$\widetilde{W}_1^n = \widetilde{W}_1^n(h) = \begin{cases} \overline{W}_1^n \text{ in the regular case,} \\ W(\mu_1 + h^4, z_n) \text{ in the singular case,} \end{cases} \tag{7.55}$$

$$\widetilde{W}_I^n = \widetilde{W}_I^n(h) = \begin{cases} \overline{W}_I^n \text{ in the regular case,} \\ W(\mu_I - h^4, z_n) \text{ in the singular case.} \end{cases} \tag{7.56}$$

Now the scheme, which is well defined whenever $I \geq 5, I$ odd, and $N \geq 3$, reads as follows (see Section 4.1):

- For $(i,n) \in \{1\} \times \{1,\dots,N-1\}$,

$$\begin{aligned} \Big(-\frac{\mu_1}{k} + \frac{\widetilde{\mathtt{A}}_1^n}{2} + \frac{\overline{\sigma}_1^n \overline{D}_2}{2h^2}\Big)\psi_1^n + \Big(-\frac{\overline{\sigma}_1^n \overline{D}_3}{8h^2}\Big)\psi_2^n + \Big(-\frac{\overline{\sigma}_1^n \overline{D}_2}{2h^2}\Big)\psi_3^n + \Big(\frac{\overline{\sigma}_1^n \overline{D}_3}{8h^2}\Big)\psi_4^n + \\ \Big(\frac{\mu_1}{k} + \frac{\widetilde{\mathtt{A}}_1^{n+1}}{2} + \frac{\overline{\sigma}_1^{n+1} \overline{D}_2}{2h^2}\Big)\psi_1^{n+1} + \Big(-\frac{\overline{\sigma}_1^{n+1} \overline{D}_3}{8h^2}\Big)\psi_2^{n+1} + \Big(-\frac{\overline{\sigma}_1^{n+1} \overline{D}_2}{2h^2}\Big)\psi_3^{n+1} + \\ \Big(\frac{\overline{\sigma}_1^{n+1} \overline{D}_3}{8h^2}\Big)\psi_4^{n+1} = \frac{\widetilde{W}_1^n + \widetilde{W}_1^{n+1}}{2}. \end{aligned} \tag{7.57}$$

- For $(i,n) \in \{2,\dots,i^*-1\} \cup \{i^*+1,\dots,I-1\} \times \{1,\dots,N-1\}$,

$$\begin{aligned} \Big(-\frac{\overline{\sigma}_i^n \overline{D}_{i-\frac{1}{2}}}{2h^2}\Big)\psi_{i-1}^n + \Big(-\frac{\mu_i}{k} + \frac{\overline{\mathtt{A}}_i^n}{2} + \frac{\overline{\sigma}_i^n \big(\overline{D}_{i-\frac{1}{2}} + \overline{D}_{i+\frac{1}{2}}\big)}{2h^2}\Big)\psi_i^n + \\ \Big(-\frac{\overline{\sigma}_i^n \overline{D}_{i+\frac{1}{2}}}{2h^2}\Big)\psi_{i+1}^n + \Big(-\frac{\overline{\sigma}_i^{n+1} \overline{D}_{i-\frac{1}{2}}}{2h^2}\Big)\psi_{i-1}^{n+1} + \\ \Big(\frac{\mu_i}{k} + \frac{\overline{\mathtt{A}}_i^{n+1}}{2} + \frac{\overline{\sigma}_i^{n+1} \big(\overline{D}_{i-\frac{1}{2}} + \overline{D}_{i+\frac{1}{2}}\big)}{2h^2}\Big)\psi_i^{n+1} + \\ \Big(-\frac{\overline{\sigma}_i^{n+1} \overline{D}_{i+\frac{1}{2}}}{2h^2}\Big)\psi_{i+1}^{n+1} = \frac{\overline{W}_i^n + \overline{W}_i^{n+1}}{2}. \end{aligned} \tag{7.58}$$

- For $(i,n) \in \{i^*\} \times \{2,\dots,N-1\}$

$$\Big(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\Big)\psi_{i^*-1}^n + \Big(\overline{\mathtt{A}}_{i^*}^n + \frac{2\overline{\sigma}_{i^*}^n}{h^2}\Big)\psi_{i^*}^n + \Big(-\frac{\overline{\sigma}_{i^*}^n}{h^2}\Big)\psi_{i^*+1}^n = \overline{W}_{i^*}^n. \tag{7.59}$$

- For $(i,n) \in \{I\} \times \{1,\dots,N-1\}$,

$$\begin{aligned} \Big(\frac{\overline{\sigma}_I^n \overline{D}_{I-2}}{8h^2}\Big)\psi_{I-3}^n + \Big(-\frac{\overline{\sigma}_I^n \overline{D}_{I-1}}{2h^2}\Big)\psi_{I-2}^n + \Big(-\frac{\overline{\sigma}_I^n \overline{D}_{I-2}}{8h^2}\Big)\psi_{I-1}^n + \\ \Big(-\frac{\mu_I}{k} + \frac{\widetilde{\mathtt{A}}_I^n}{2} + \frac{\overline{\sigma}_I^n \overline{D}_{I-1}}{2h^2}\Big)\psi_I^n + \Big(\frac{\overline{\sigma}_I^{n+1} \overline{D}_{I-2}}{8h^2}\Big)\psi_{I-3}^{n+1} + \Big(-\frac{\overline{\sigma}_I^{n+1} \overline{D}_{I-1}}{2h^2}\Big)\psi_{I-2}^{n+1} + \\ \Big(-\frac{\overline{\sigma}_I^{n+1} \overline{D}_{I-2}}{8h^2}\Big)\psi_{I-1}^{n+1} + \Big(\frac{\mu_I}{k} + \frac{\widetilde{\mathtt{A}}_I^{n+1}}{2} + \frac{\overline{\sigma}_I^{n+1} \overline{D}_{I-1}}{2h^2}\Big)\psi_I^{n+1} = \frac{\widetilde{W}_I^n + \widetilde{W}_I^{n+1}}{2}. \end{aligned} \tag{7.60}$$

- For $(i,n) \in \{i^*,\dots,I\} \times \{1\}$,

$$\psi_i^1 = \overline{f}_i. \tag{7.61}$$

- For $(i,n) \in \{1,\dots,i^*\} \times \{N\}$,
$$\psi_i^N = \overline{g}_i. \tag{7.62}$$

The order 2 of the scheme was evinced in Subsection 3.1.2 for the regular case. The fourth power h^4 in Equations *(7.53)–(7.56)* has been chosen in order to preserve the order 2 in the tests conducted for singular cases; numerical results are shown in the next Subsection 7.4.2.

7.4.2. Numerical experiments with the core scheme

Only numerical tests for the singular case are carried out, because the scheme has already been tested for the regular case in Subsection 3.1.2. In the tables below, $E_{\text{abs}} = \max|\psi_{\text{grid}} - \psi|$, where ψ_{grid} is representing the approximate solution and the maximum is taken over the set of all nodes.

Let us take

$$Z_{\text{ini}} = 0,\ Z_{\text{fin}} = 1, \tag{7.63}$$
$$\sigma(\mu,z) = 1 + \sin(12\mu z)\cos(12\mu z), \tag{7.64}$$
$$\text{and } \mathtt{A}(\mu,z) = |\sin(12\mu z)| + \frac{\sigma(\mu,z)}{1-\mu^2}. \tag{7.65}$$

When an exact solution is given, W is computed so that Equation *(7.44)* be satisfied. The following numerical experiments have been conducted:

1. Let us consider the exact solution $\psi(\mu,z) = \ln(2+\mu^2+z^3)$. Then, $f(\mu) = \ln(2+\mu^2)$, $g(\mu) = \ln(3+\mu^3)$. Notice that W is unbounded and not defined, in an essential way, at $|\mu| = 1$. Results are collected in Table 7.4.1.

(I,N)	E_{abs}	order
$(11,10)$	3.04×10^{-3}	
$(33,29)$	3.37×10^{-4}	$\frac{2\ln(\frac{30.4}{3.37})}{\ln(10)} = 1.91$
$(101,91)$	3.61×10^{-5}	1.94
$(321,281)$	3.58×10^{-6}	2.01
$(1001,901)$	3.67×10^{-7}	1.98

Table 7.4.1: Maximum error and order.

Since

$$\mathcal{H}(\pm 1, z) = 1 \pm \sin(12z)\cos(12z) \tag{7.66}$$

and W satisfies *(7.48)* with

$$\zeta_{\pm}(z) = \big(1 \pm \sin(12z)\cos(12z)\big)\ln(3+z^3), \tag{7.67}$$

we should have $\psi(\pm 1, z) = \frac{\zeta_{\pm}(z)}{\mathcal{H}(\pm 1, z)} = \ln(3 + z^3)$, which is obviously true.

2. Let us consider the exact solution $\psi(\mu, z) = (1 - \mu^2)\ln(2 + z^3)$. Then, $f(\mu) = (1 - \mu^2)\ln 2$, $g(\mu) = (1 - \mu^2)\ln 3$. Notice that W is bounded and can be defined in a continuous way at $|\mu| = 1$. Results are shown in Table 7.4.2.

(I, N)	E_{abs}	order
$(11, 10)$	8.15×10^{-3}	
$(33, 29)$	6.32×10^{-4}	$\frac{2\ln(\frac{81.5}{6.32})}{\ln(10)} = 2.22$
$(101, 91)$	6.76×10^{-5}	1.94
$(321, 281)$	6.71×10^{-6}	2.01
$(1001, 901)$	6.90×10^{-7}	1.98

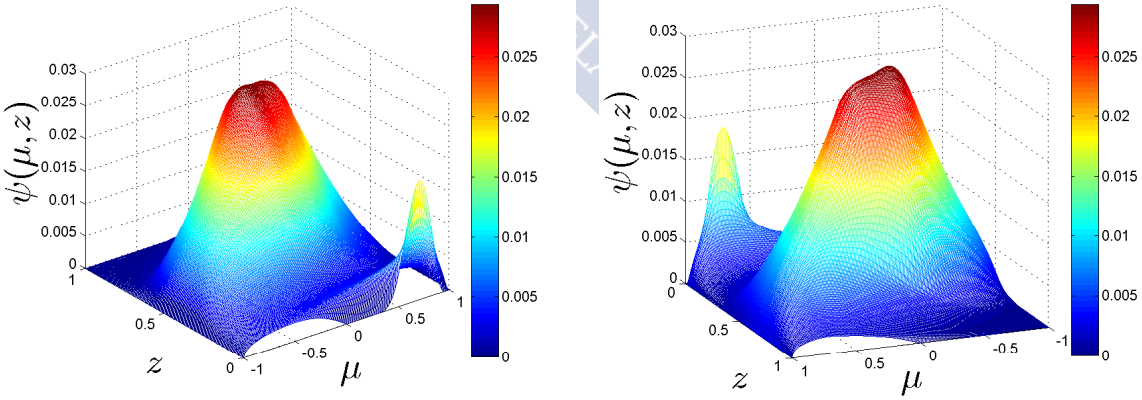
Table 7.4.2: Maximum error and order.

In this case W satisfies *(7.47)* and so we should have $\psi(\pm 1, z) = 0$, which is again obviously true.

3. $f(\mu) = \exp\big[-1/\big(\mu^2(1 - \mu^2)\big)\big]$, $g(\mu) = 0$

$$W(\mu, z) = \begin{cases} |2\mu z| & \text{if } \|(\mu, z) - (0, 0.5)\|_2 \leq 0.3 \\ 0 & \text{otherwise.} \end{cases}$$

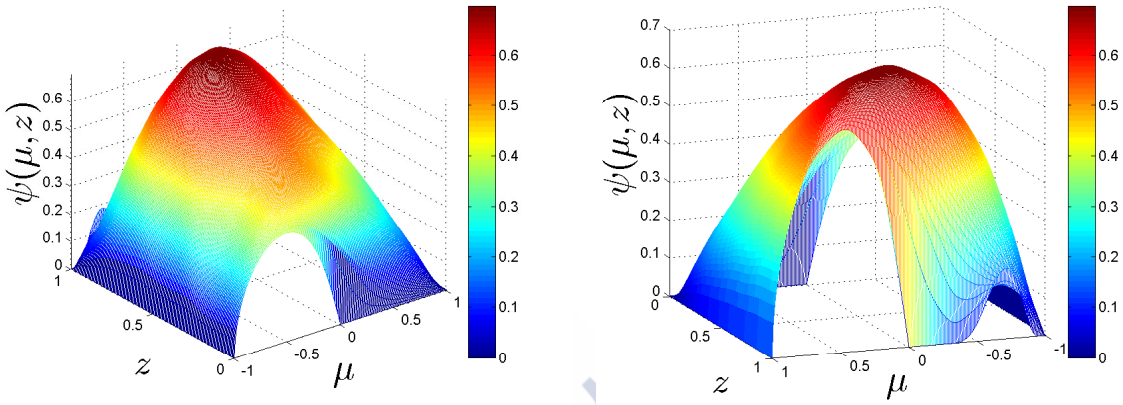
In this case, the exact solution is not known. Notice that W is discontinuous. However, the observed solution ψ, plotted in Figure 7.1, is continuous due to the effect of the diffusion. Notice that W satisfies *(7.47)*, and observe that $\psi(\pm 1, z) = 0$.



(a) Front. (b) Back.

Figure 7.1: Approximate solution for $I = 201$, $N = 200$.

4. $f(\mu) = 0,\ g(\mu) = 10\exp\big[-1/\big(\mu^2(1-\mu^2)\big)\big],\ W(\mu, z) = \frac{1+z}{\sqrt{1-\mu^2}}$.
   This is another example with unknown exact solution. Notice that the observed solution ψ, plotted in Figure 7.2, is zero at $|\mu| = 1$, despite being the source W unbounded. Once more, this is because W satisfies condition *(7.47)*.



(a) Front. (b) Back.

Figure 7.2: Approximate solution for $I = 201$, $N = 200$.

7.5. Numerical results for the full problem

The solution of problem *(7.1)–(7.4)* ψ is approximated from Equation *(7.27)* via a truncation of the Fourier series, that is to say,

$$\psi(\mu, z, \theta) \approx \psi_0(\mu, z) + 2\sum_{k=1}^{K} \big\{\mathrm{Re}[\psi_k(\mu, z)]\cos k\theta - \mathrm{Im}[\psi_k(\mu, z)]\sin k\theta\big\} \tag{7.68}$$

and each ψ_k, $k = 0, \ldots, K$ is computed as the solution to the problem *(7.35)–(7.37)*. The case $K = 0$ must be understood as $\psi(\mu, z, \theta) \approx \psi_0(\mu, z)$.

When W, f and g do not depend on θ, then the solution ψ is θ-independent as well and the problem solved is actually the one solved in Chapter 3; taking $K = 0$ is enough in this case.

When $\psi(\mu, z, \theta) = \psi_0(\mu, z) + \psi_{\cos}(\mu, z)\cos(m\theta) + \psi_{\sin}(\mu, z)\sin(n\theta)$, with $m, n \in \mathbb{N}$, it is enough to solve for $k = 0, m, n$ in order to have exactness with respect to the variable θ. These simple cases will be avoided in the tests below.

Notice that Equation *(7.68)* allows computing approximations of the values $\psi(\mu_i, z_n, \theta)$ on the grid nodes (μ_i, z_n), for every $\theta \in [0, 2\pi)$. In subsequent tables, we will use the notation $E_{\mathrm{abs}}^{\mathrm{full}} = \max|\psi_{\mathrm{grid}} - \psi|$, where ψ_{grid} is representing the approximate solution and the maximum is taken over the set

$$\{(\mu_i, z_n, \theta_j) :\ 1 \le i \le I,\ 1 \le n \le N,\ 1 \le j \le J\}, \tag{7.69}$$

being $\theta_1 = 0 < \theta_2 < \ldots < \theta_{J-1} < \theta_J = 2\pi$ a uniform mesh of $[0, 2\pi]$ with $J = 100$.

We will always take $Z_{\text{ini}} = 0$, $Z_{\text{fin}} = 1$.

7.5.1. Test #1

Let us consider

$$\alpha(\mu, z) = |\sin(12\mu z)|, \tag{7.70}$$
$$\sigma(\mu, z) = 1 + \sin(12\mu z)\cos(12\mu z). \tag{7.71}$$

The function $\psi(\mu, z, \theta) = \ln\big(1+\sin(3\mu z)\big) + (1-\mu^2)\mathrm{e}^{\sin\theta}$ is the exact solution of problem *(7.1)–(7.4)* if

$$f(\mu, \theta) = \ln(2) + (1-\mu^2)\mathrm{e}^{\sin\theta}, \tag{7.72}$$
$$g(\mu, \theta) = \ln\big(2+\sin(3\mu z)\big)(1-\mu^2)\mathrm{e}^{\sin\theta}. \tag{7.73}$$

and

$$\begin{aligned} W(\mu, z, \theta) = {} & \frac{3\mu^2\cos(3\mu z)}{2+\sin(3\mu z)} + \alpha(\mu, z)\psi(\mu, z, \theta) - \\ & \sigma(\mu, z)\Bigg\{ -2\mu\left(\frac{3z\cos(3\mu z)}{2+\sin(3\mu z)} - 2\mu\mathrm{e}^{\sin\theta}\right) + \\ & (1-\mu^2)\left(\frac{-18z^2\sin(3\mu z) - 9z^2}{\big(2+\sin(3\mu z)\big)^2} - 2\mathrm{e}^{\sin\theta}\right) + \\ & (-\sin\theta + (\cos\theta)^2)\mathrm{e}^{\sin\theta}\Bigg\}. \end{aligned} \tag{7.74}$$

Since the Fourier series of $\mathrm{e}^{\sin\theta}$ is infinite, there is no expectancy that Equation *(7.59)* be exact for any value of K. However, the high regularity of this function make the series converge fast, and so a small value of K is enough to capture the solution. Tables 7.5.1, 7.5.2 and 7.5.3 show the results obtained for $K = 5, 6, 7$. One can compare the results for successive values of K and consider that the solution is good when there is no significant difference among them, but the criterion for not going beyond $K = 7$ in this case was simply that, for the meshes we are using, this value of K preserves for the full problem the order 2 of the core scheme; in other words, it is not worth taking $K > 7$ is negligible when compared with the error due to discretization of the (μ, z)-domain.

(I, N)	$E^{\text{full}}_{\text{abs}}$	order
$(11, 10)$	1.24×10^{-1}	
$(33, 29)$	3.74×10^{-3}	$\frac{2\ln(\frac{124}{3.74})}{\ln(10)} = 3.04$
$(101, 91)$	4.04×10^{-4}	1.93
$(321, 281)$	8.17×10^{-5}	1.38
$(1001, 901)$	5.11×10^{-5}	0.41

Table 7.5.1: Numerical results for test #1 with $K = 5$.

(I, N)	$E_{\text{abs}}^{\text{full}}$	order
$(11, 10)$	1.24×10^{-1}	
$(33, 29)$	3.74×10^{-3}	$\frac{2\ln(\frac{124}{3.74})}{\ln(10)} = 3.04$
$(101, 91)$	3.77×10^{-4}	1.99
$(321, 281)$	4.09×10^{-5}	1.93
$(1001, 901)$	6.99×10^{-6}	1.53

Table 7.5.2: Numerical results for test #1 with $K = 6$.

(I, N)	$E_{\text{abs}}^{\text{full}}$	order
$(11, 10)$	1.24×10^{-1}	
$(33, 29)$	3.74×10^{-3}	$\frac{2\ln(\frac{124}{3.74})}{\ln(10)} = 3.04$
$(101, 91)$	3.79×10^{-4}	1.99
$(321, 281)$	3.92×10^{-5}	1.97
$(1001, 901)$	4.18×10^{-6}	1.94

Table 7.5.3: Numerical results for test #1 with $K = 7$.

7.5.2. Test #2

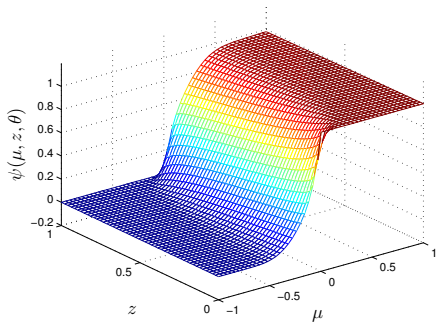
In this test, we take the following data functions:

$$\alpha(\mu, z) = 0.02, \tag{7.75}$$
$$\sigma(\mu, z) = 0.01, \tag{7.76}$$
$$f(\mu, \theta) = 1, \tag{7.77}$$
$$g(\mu, \theta) = (1 + \mu)\sin(3\theta), \tag{7.78}$$
$$W(\mu, z, \theta) = 0. \tag{7.79}$$

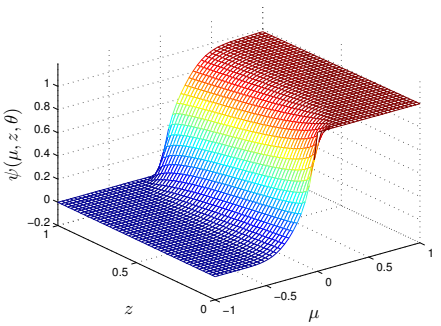
In this case, both the Fourier coefficients f_k, g_k and W_k and the solution ψ_k of problem *(7.35)–(7.37)* are zero, when $K \neq 0$ and $K \neq 3$. Consequently the numerical method is exact with respect to θ if we take $K = 3$.

This example has been chosen in order to show that the approximate solution can be singular at $(\mu, z, \theta) \in \{(0, Z_{\text{ini}}, \cdot), (0, Z_{\text{fin}}, \cdot)\} = \{(0, 0, \cdot), (0, 1, \cdot)\}$ even when all data functions are of class C^∞. To specify, $\frac{\partial\psi}{\partial\mu}(0, 0, \cdot)$ and $\frac{\partial\psi}{\partial\mu}(0, 1, \cdot)$ do not exist and $|\frac{\partial\psi}{\partial z}(0, z, \cdot)| \to \infty$ when $z \to 0$ and also when $z \to 1$.

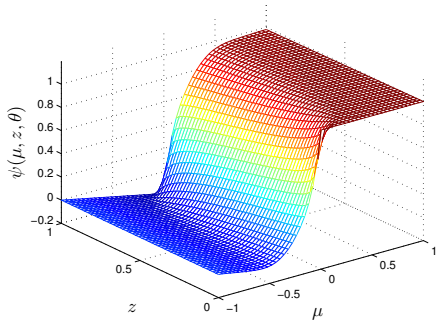
These comments are apparent when one looks at the figures below:



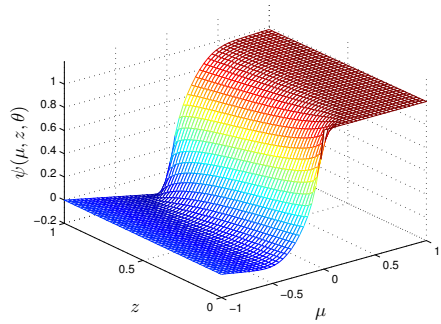
(a) At $\theta = 0.00$.



(b) At $\theta = 2.03$.

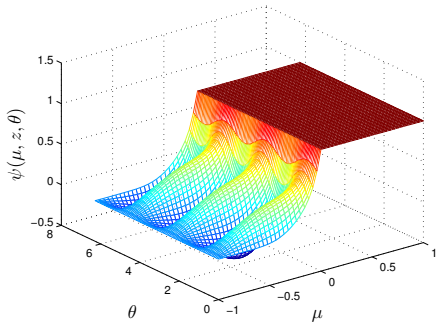


(c) At $\theta = 4.13$.

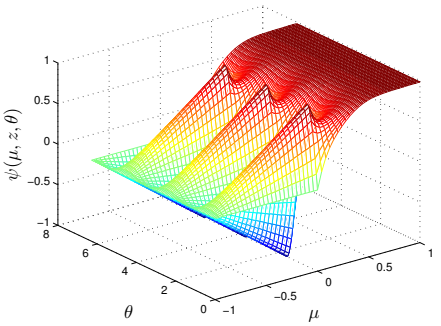


(d) At $\theta = 6.28$.

Figure 7.3: Approximate solution of test #2 obtained with the mesh $(I, N) = (51, 50)$ at different values of θ. Figures 7.3(a) and 7.3(d) are the same because of the periodic condition *(7.4)*.



(a) At $z = Z_{\text{ini}} = 0$.



(b) At $z = Z_{\text{fin}} = 1$.

Figure 7.4: Approximate solution of test #2.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Conclusions and future work

Conclusions

In this thesis we have presented new numerical methods, called "direct method"and "iterative method" for solving the FPE with a source term W in the 1D slab. Coefficients and the source term W, in general, could depend on the three variables (z, μ, θ), the case which is not considered in the papers, for instance in [21], [32], [33]. In the case of θ-dependence, Fourier techniques are used to divide the problem into a set of θ-independent problems, whose absorption coefficient α turns into singular. In order to solve the problems it is required to modify the odd scheme.

Several plots and other numerical results that support the validity of the direct method consisting of the even and the odd schemes, have been shown. All numerical results carried out with the above-mentioned schemes verify that the second one, the odd scheme is preferable. They confirm convergence of order 2 in both variables z, μ. Regarding the convergence with respect to θ, we cannot say exact order of convergence because we have not employed mesh in θ variable. It is provided by the convergence velocity of Fourier series.

As for the iterative method, which most scientists have suggested in their papers solving such forward-backward problems as the FPEs (see [21], [32], [33]), does not give better results from the viewpoint of computational time than the direct one unless very fine is required. Numerical tests carried out with both direct and iterative methods show that the direct method is best option for the problems considered in this thesis and above-mentioned papers.

Future work

Throughout this thesis we have considered time and energy-independent problem. Gorbikov and Melnikov in [15], following first order finite difference scheme, solved time and energy-dependent Fokker-Planck equation and obtained some results. Being motivated by the work [15], we are going to investigate the problem *(1)–(3)* when the angular flux ψ is a function of $(x, t, \mu, \theta, \epsilon)$, which is general case. The methods used in the thesis can also be used for the time and energy-dependent problem after performing discretizations in these variables. The results are supposed to show convergence of order 2, if an appropriate finite difference scheme is employed. As numerical results are not presented in [15], comparisons could not be possible.

From the viewpoint of the numerical methods, it could also be possible to employ approriate numerical method for the PDE *(1)* on surfaces. Some finite element methods have been studied for the Laplace-Beltrami operator *(7.5)* in [8], [9] on various surfaces. And morever finite element methods for surface PDEs has been investigated by Dziuk and Elliott in [10].

Resumen en castellano
Summary in Spanish

La tesis se centra en la resolución numérica, mediante diferencias finitas y técnicas de Fourier, de dos ecuaciones cinéticas de tipo Fokker-Planck. Puesto que la denominación "ecuación de Fokker-Planck" (EFP) alberga distintas interpretaciones, es preciso decir que este trabajo trata de aquella que proviene del ámbito de la ingeniería nuclear, donde la incógnita representa la densidad de flujo angular de partículas, las cuales son generalmente partículas cargadas tales como electrones o iones pesados.

El principal interés de la EFP radica en que su solución es, bajo ciertas hipótesis, una buena aproximación de la ecuación de transporte de Boltzmann, a la vez que su resolución numérica es menos costosa que esta última. Por citar alguna aplicación, la EFP se ha utilizado en el marco de la radioterapia externa como modelo para el transporte de electrones a través del cuerpo humano y en el de las imágenes tomográficas como modelo de la propagación de la luz en tejidos biológicos.

Los casos estudiados en la presente tesis son particularizaciones del problema siguiente: dado un dominio espacial (no vacío, abierto, acotado y convexo) $\Omega \subset \mathbb{R}^3$ y un intervalo de tiempo $[T_{\text{ini}}, T_{\text{fin}}]$, hallar la función

$$\begin{array}{rcl} \psi : \overline{\Omega} \times [T_{\text{ini}}, T_{\text{fin}}] \times [-1,1] \times [0,2\pi] \times (0,\infty) & \longrightarrow & \mathbb{R} \\ (\boldsymbol{x}, t, \mu, \theta, \epsilon) & \longrightarrow & \psi(\boldsymbol{x}, t, \mu, \theta, \epsilon) \end{array} \tag{8.1}$$

que satisfaga tanto la EFP

$$\frac{1}{c}\frac{\partial \psi}{\partial t} + \boldsymbol{\omega} \cdot \nabla \psi + \alpha \psi = \sigma \left\{ \frac{\partial}{\partial \mu} \left[\left(1-\mu^2\right) \frac{\partial \psi}{\partial \mu} \right] + \frac{1}{1-\mu^2} \frac{\partial^2 \psi}{\partial \theta^2} \right\} + \frac{\partial (S_M \psi)}{\partial \epsilon} + W \tag{8.2}$$

como ciertas condiciones de cierre adecuadas, como pueden ser por ejemplo

$$\psi_{|\{t=T_{\text{ini}}\}} = \widehat{\psi}, \text{ con } \widehat{\psi} = \widehat{\psi}(\boldsymbol{x}, \mu, \theta, \epsilon) \text{ dada}, \tag{8.3}$$

$$\psi_{|\{(\boldsymbol{x},\boldsymbol{\omega}) \in \partial\Omega \times S^2 :\, \boldsymbol{\omega}\cdot\boldsymbol{n}(\boldsymbol{x}) < 0\}} = \mathscr{F}, \text{ con } \mathscr{F} = \mathscr{F}(\boldsymbol{x}, t, \mu, \theta, \epsilon) \text{ dada}, \tag{8.4}$$

$$\psi_{|\{\theta=0\}} = \psi_{|\{\theta=2\pi\}}, \quad \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=0\}} = \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta=2\pi\}}, \tag{8.5}$$

$$\psi_{|\{\epsilon=\infty\}} = 0. \tag{8.6}$$

Es interesante observar el hecho de que el problema *(8.2)–(8.6)* no necesita ni admite condiciones de cierre en $\mu \in \{-1, 1\}$, lo que puede relacionarse con la degeneración del coeficiente de difusión interno $1-\mu^2$ en esos puntos.

El significado de las notaciones empleadas es el siguiente:

- $\psi = \psi(\boldsymbol{x}, t, \mu, \theta, \epsilon)$: densidad de flujo angular de partículas.
- $\boldsymbol{x} = (x_1, x_2, x_3) \in \overline{\Omega} \subset \mathbb{R}^3$ representa un punto del dominio espacial.
- $t \in [T_{\text{ini}}, T_{\text{fin}}]$ representa el tiempo.
- c: velocidad de partículas en el medio.
- S^2: esfera unidad en $\mathbb{R}^3$ (conjunto de vectores de $\mathbb{R}^3$ de módulo 1).
- $\boldsymbol{\omega} \in S^2$ representa la dirección en la que se mueve la partícula. Se tiene

$$\boldsymbol{\omega} = \boldsymbol{\omega}(\varphi, \theta) = (\sin\varphi\cos\theta, \sin\varphi\sin\theta, \cos\varphi), \tag{8.7}$$

  con $\varphi \in [0, \pi]$ el ángulo polar y $\theta \in [0, 2\pi)$ el ángulo acimutal en el sistema estándar de coordenadas esféricas centradas en la posición $\boldsymbol{x}$ de la partícula.
- ∇ se usa en este caso para referirse al gradiente con respecto a las tres variables espaciales. Es decir, $(\nabla\psi)_i = \frac{\partial\psi}{\partial x_i}$ para $i \in \{1, 2, 3\}$.
- $\mu = \omega_3 = \cos\varphi \in [-1, 1]$. Nótese que μ determina de forma única el ángulo polar φ.
- $\epsilon \in (0, \infty)$ es la energía de la partícula.

En la situación más general, las funciones dato α, σ y W que aparecen en la EFP *(8.2)* pueden depender de $(\boldsymbol{x}, t, \mu, \theta, \epsilon)$. La función W modela una fuente de (densidad de flujo angular de) partículas allí donde es positiva y un sumidero donde es negativa; con independencia del signo, suele llamarse "fuente". Los términos que van acompañados de las funciones $\alpha \geq 0$ y $\sigma > 0$ modelan, respectivamente, los fenómenos de absorción y de "cambios de dirección por colisiones" (*scattering*). La función $S_M = S_M(\boldsymbol{x}, \epsilon)$ es conocida como la "potencia de parada" (*stopping power*), y representa una media de la energía perdida en una unidad de "camino recorrido" (*path length*); la idea es que las colisiones inelásticas entre los electrones tienen lugar con tanta frecuencia que, como aproximación, puede considerarse que experimentan de forma continua una pérdida energética fija por unidad de camino recorrido.

En la EFP *(8.2)*, las ganancias por *scattering* hacia la dirección $\boldsymbol{\omega}$ de partículas provenientes de otras direcciones se representan por el término

$$\sigma \left\{ \frac{\partial}{\partial\mu} \left[\left(1 - \mu^2\right) \frac{\partial\psi}{\partial\mu} \right] + \frac{1}{1-\mu^2} \frac{\partial^2\psi}{\partial\theta^2} \right\}. \tag{8.8}$$

Se hace notar que el operador $\frac{\partial}{\partial\mu}\left[\left(1-\mu^2\right)\frac{\partial\psi}{\partial\mu}\right] + \frac{1}{1-\mu^2}\frac{\partial^2\psi}{\partial\theta^2}$ es el laplaciano de ψ sobre la esfera S^2, por lo que se está modelando, en las variables angulares, la difusión típica. Este hecho conecta la EFP con las EDP sobre superficies, sobre las que existe abundante bibliografía que se podría explorar; en esta memoria no se ha seguido ese camino.

En el caso estacionario, independiente de la energía, y en geometría de "laja" (*slab*)

unidimensional $\Omega = \Omega^* \times (Z_{\text{ini}}, Z_{\text{fin}})$, con $\Omega^* \subset \mathbb{R}^2$, la EFP queda reducida a

$$\mu \frac{\partial \psi}{\partial z} + \alpha \psi = \sigma \left\{ \frac{\partial}{\partial \mu} \left[\left(1 - \mu^2\right) \frac{\partial \psi}{\partial \mu} \right] + \frac{1}{1 - \mu^2} \frac{\partial^2 \psi}{\partial \theta^2} \right\} + W, \tag{8.9}$$

donde $z = x_3$, $\psi : [-1, 1] \times [Z_{\text{ini}}, Z_{\text{fin}}] \times [0, 2\pi] \to \psi(z, \mu, \theta) \in \mathbb{R}$, y las condiciones de cierre se adaptan convenientemente:

$$\psi_{|\{\mu \in (0,1],\ z = Z_{\text{ini}}\}} = f, \text{ con } f = f(\mu, \theta) \text{ dada}, \tag{8.10}$$

$$\psi_{|\{\mu \in [-1,0),\ z = Z_{\text{fin}}\}} = g, \text{ con } g = g(\mu, \theta) \text{ dada}, \tag{8.11}$$

$$\psi_{|\{\theta = 0\}} = \psi_{|\{\theta = 2\pi\}}, \quad \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta = 0\}} = \left(\frac{\partial \psi}{\partial \theta}\right)_{|\{\theta = 2\pi\}}. \tag{8.12}$$

En el capítulo 1 se explica cómo se obtiene la ecuación *(8.9)* a partir de la ecuación *(8.2)* bajo las hipótesis simplificadoras mencionadas en el párrafo anterior. También se explica una forma de reducir el coste computacional expresando ψ como una serie de Fourier con respecto a la variable θ, que, de acuerdo con la condición *(8.12)*, es la variable de periodicidad. Esta "técnica de Fourier" consiste en resolver varios problemas independientes de θ, uno para cada coeficiente de Fourier, con la consecuencia de que se evita tener que hacer una discretización en esa variable; funciona de forma directa en geometría de laja unidimensional cuando los coeficientes de Fourier decaen rápidamente y se usa en el capítulo 7. Hasta ese capítulo 7, los esfuerzos se dirigen a resolver el problema independiente de θ, es decir,

$$\mu \frac{\partial \psi}{\partial z} + \alpha \psi = \sigma \frac{\partial}{\partial \mu} \left[\left(1 - \mu^2\right) \frac{\partial \psi}{\partial \mu} \right] + W, \tag{8.13}$$

con las condiciones de cierre

$$\psi_{|\{\mu \in (0,1],\ z = Z_{\text{ini}}\}} = f, \text{ con } f = f(\mu) \text{ dada}, \tag{8.14}$$

$$\psi_{|\{\mu \in [-1,0),\ z = Z_{\text{fin}}\}} = g, \text{ con } g = g(\mu) \text{ dada}. \tag{8.15}$$

Este problema se categoriza dentro de las llamadas "ecuaciones de difusión de dos vías" (*two-way diffusion equations*) o también "ecuaciones de difusión de ida y vuelta") (*forward-backward diffusion equations*).

La fuente W en la ecuación *(8.13)* suele aparecer en la literatura igual a 0. La inclusión de una fuente que pueda ser no nula ofrece al menos dos ventajas:

- Por un lado, permite diseñar problemas con solución exacta conocida: basta elegir libremente una función ψ que pueda ser derivada un número suficiente de veces, por ejemplo que sea de clase C^∞, así como las funciones α y σ, y luego ajustar las funciones W, f y g para que ψ sea la solución del problema. Disponer de este tipo de ejemplos, con solución exacta conocida y regular, permite comprobar experimentalmente que el orden de convergencia del esquema es realmente igual a 2.

- Por otro lado, tanto el problema evolutivo como el problema dependiente de la energía pueden discretizarse dando lugar a problemas "estacionarios" como el problema *(8.13)*–*(8.15)*, con fuentes W que son no nulas.

Cabe decir también que el problema *(8.13)–(8.15)* es nuclear, en el sentido de que los problemas más generales anteriormente descritos pueden ser resueltos si se sabe resolver este. Es conveniente disponer, pues, de un método robusto para resolver el problema *(8.13)–(8.15)*. Cuando se trata de diseñar métodos numéricos, es importante tener en mente propiedades conocidas de la solución exacta cuando ello es posible, y una de las propiedades de mayor importancia en este contexto es la regularidad en el sentido clásico de las funciones de clase C^k. Hablando desde un punto de vista puramente teórico, no existen hasta la fecha resultados de regularidad clásica, pero sí se tienen evidencias numéricas de que la solución puede no ser regular aun cuando las funciones dato sean constantes.

Por ello, se dedica un pequeño capítulo 2 al estudio de la regularidad de la solución del problema *(8.13)–(8.15)* cuando α es una constante positiva y $\sigma \equiv W \equiv 0$, caso en el que la EDP se trivializa en una EDO cuya solución exacta puede calcularse explícitamente. Este ejemplo particular nos previene sobre la posible falta de regularidad de ψ en situaciones más complejas, lo cual, efectivamente y como se anunciaba en el párrafo anterior, se observa más adelante en los resultados numéricos a pesar del efecto regularizante de la difusividad (excepto en este capítulo 2, siempre se considera el caso, físicamente más relevante, en el que $\sigma > 0$).

En el capítulo 3 se propone un método basado en diferencias finitas para resolver el problema *(8.13)–(8.15)*. Para diseñarlo, se tiene en cuenta que, aunque pueda no ser regular, la solución ψ es normalmente continua; concretamente, la falta de regularidad se observa en alguna de las derivadas de primer orden, porque el hecho de que la difusividad σ sea positiva hace que ψ sea continua en una amplia gama de situaciones. Además, la similitud con el problema del calor unidimensional lleva a pensar en la conveniencia de emplear un método implícito para que sea robusto. Comoquiera que este problema es "de valor inicial" (PVI) cuando $\mu > 0$ pero "de valor final" (PVF) cuando $\mu < 0$, y que un método implícito debe escribirse "hacia atrás" (*backward*) para los PVI pero "hacia delante" (*forward*) para los PVF, es inmediato pensar en un esquema de tipo Crank-Nicolson como un candidato idóneo, al menos desde el punto de vista de la escritura. En efecto, al tener ese esquema una parte explícita y una implícita con el mismo peso $\frac{1}{2}$, puede ser empleado tanto en la parte con $\mu > 0$ como en la parte con $\mu < 0$ sin necesidad de modificar su escritura. Se usa por lo tanto la idea de Crank-Nicolson para la discretización con respecto a z, que en principio promete orden 2 con respecto a esa variable, lo cual a su vez lleva a usar discretizaciones de orden 2 para la variable μ con objeto de diseñar un método que sea de orden 2 con respecto a ambas variables. La discretización de orden 2 con respecto a μ debe hacerse con cuidado en $\mu \in \{-1, 1\}$, donde se ha optado por fórmulas descentradas que aprovechen la degeneración de $1 - \mu^2$ en esos puntos. Procediendo de esta forma, se diseñan dos esquemas para mallas uniformes de I μ-nodos y N z-nodos: el esquema par, válido cuando I es par, y el esquema impar, válido cuando I es impar. Mediante experimentos numéricos se muestra que en efecto ambos esquemas son de orden 2 con respecto a z y a μ, se observa la superioridad del esquema impar sobre el par en materia de estabilidad numérica sobre un ejemplo concreto y se llama la atención sobre la posibilidad de emplear el conocimiento del orden para sacar conclusiones sobre la posible falta de regularidad de ψ. Con respecto al último asunto, la idea es que el orden de convergencia solamente se aprecia cuando la solución es lo bastante regular y, por lo tanto, cuando al emplear un método de orden 2 se observa para un determinado problema un orden claramente inferior a 2 en los experimentos numéricos, hay que pensar que se tienen indicaciones de carencias en la regularidad de la solución. Para calcular el orden cuando no se conoce la solución exacta, se introduce el concepto de orden estrella.

Los dos esquemas descritos en el capítulo 3 pueden ser resueltos mediante un método directo o mediante un método iterativo. En el primero, las aproximaciones en todos los puntos de la malla se obtienen de forma directa al resolver un único sistema lineal de orden $(I \times N) \times (I \times N)$, procediendo, para entendernos, como se hace al resolver el problema de Poisson bidimensional en un rectángulo. Precisamente es el método directo el que se emplea en el capítulo 3, mientras que se pospone la descripción del iterativo hasta el capítulo 5.

Sea $h = \frac{2}{I-1}$ la distancia entre los μ-nodos $-1 = \mu_1 < \ldots < \mu_I = 1$ y $k = \frac{Z_{\text{fin}} - Z_{\text{ini}}}{N-1}$ la distancia entre los z-nodos $Z_{\text{ini}} = z_1 < \ldots < z_N = Z_{\text{fin}}$. Supongamos que I (el número de μ-nodos) es impar, con lo cual $\mu_{i^\star} = 0$ si $i^\star = \frac{I+1}{2}$. Considérense también las notaciones siguientes: $D(\mu) = 1 - \mu^2$, $\overline{D}_i = D(\mu_i)$, $\overline{D}_{i\pm\frac{1}{2}} = D(\mu_i \pm \frac{h}{2})$; $\overline{\alpha}_i^n = \alpha(\mu_i, z_n)$, $\overline{\sigma}_i^n = \sigma(\mu_i, z_n)$, $\overline{W}_i^n = W(\mu_i, z_n)$; $\overline{f}_i = f(\mu_i)$, $\overline{g}_i = g(\mu_i)$; $\psi_i^n \approx \psi(\mu_i, z_n)$, donde debe entenderse que ψ es la solución exacta del problema *(8.13)–(8.15)*. Puede entenderse ahora la siguiente escritura del esquema impar, el cual proporciona $I \times N$ ecuaciones para las $I \times N$ incógnitas ψ_i^n:

- Para $(i,n) \in \{1\} \times \{1, \ldots, N-1\}$,

$$\begin{aligned}\left(-\frac{\mu_1}{k} + \frac{\overline{\alpha}_1^n}{2} + \frac{\overline{\sigma}_1^n \overline{D}_2}{2h^2}\right)\psi_1^n + \left(-\frac{\overline{\sigma}_1^n \overline{D}_3}{8h^2}\right)\psi_2^n + \left(-\frac{\overline{\sigma}_1^n \overline{D}_2}{2h^2}\right)\psi_3^n + \left(\frac{\overline{\sigma}_1^n \overline{D}_3}{8h^2}\right)\psi_4^n + \\ \left(\frac{\mu_1}{k} + \frac{\overline{\alpha}_1^{n+1}}{2} + \frac{\overline{\sigma}_1^{n+1} \overline{D}_2}{2h^2}\right)\psi_1^{n+1} + \left(-\frac{\overline{\sigma}_1^{n+1} \overline{D}_3}{8h^2}\right)\psi_2^{n+1} + \left(-\frac{\overline{\sigma}_1^{n+1} \overline{D}_2}{2h^2}\right)\psi_3^{n+1} + \\ \left(\frac{\overline{\sigma}_1^{n+1} \overline{D}_3}{8h^2}\right)\psi_4^{n+1} = \frac{\overline{W}_1^n + \overline{W}_1^{n+1}}{2}. \qquad (8.16)\end{aligned}$$

- Para $(i,n) \in (\{2, \ldots, i^\star - 1\} \cup \{i^\star + 1, \ldots, I-1\}) \times \{1, \ldots, N-1\}$,

$$\begin{aligned}\left(-\frac{\overline{\sigma}_i^n \overline{D}_{i-\frac{1}{2}}}{2h^2}\right)\psi_{i-1}^n + \left(-\frac{\mu_i}{k} + \frac{\overline{\alpha}_i^n}{2} + \frac{\overline{\sigma}_i^n \left(\overline{D}_{i-\frac{1}{2}} + \overline{D}_{i+\frac{1}{2}}\right)}{2h^2}\right)\psi_i^n + \\ \left(-\frac{\overline{\sigma}_i^n \overline{D}_{i+\frac{1}{2}}}{2h^2}\right)\psi_{i+1}^n + \left(-\frac{\overline{\sigma}_i^{n+1} \overline{D}_{i-\frac{1}{2}}}{2h^2}\right)\psi_{i+1}^{n+1} + \\ \left(\frac{\mu_i}{k} + \frac{\overline{\alpha}_i^{n+1}}{2} + \frac{\overline{\sigma}_i^{n+1} \left(\overline{D}_{i-\frac{1}{2}} + \overline{D}_{i+\frac{1}{2}}\right)}{2h^2}\right)\psi_i^{n+1} + \\ \left(-\frac{\overline{\sigma}_i^{n+1} \overline{D}_{i+\frac{1}{2}}}{2h^2}\right)\psi_{i+1}^{n+1} = \frac{\overline{W}_i^n + \overline{W}_i^{n+1}}{2}. \qquad (8.17)\end{aligned}$$

- Para $(i,n) \in \{i^\star\} \times \{2, \ldots, N-1\}$,

$$\left(-\frac{\overline{\sigma}_{i^\star}^n}{h^2}\right)\psi_{i^\star-1}^n + \left(\overline{\alpha}_{i^\star}^n + \frac{2\overline{\sigma}_{i^\star}^n}{h^2}\right)\psi_{i^\star}^n + \left(-\frac{\overline{\sigma}_{i^\star}^n}{h^2}\right)\psi_{i^\star+1}^n = \overline{W}_{i^\star}^n. \qquad (8.18)$$

- Para $(i,n) \in \{I\} \times \{1, \ldots, N-1\}$,

$$\left(\frac{\overline{\sigma}_I^n \overline{D}_{I-2}}{8h^2}\right)\psi_{I-3}^n + \left(-\frac{\overline{\sigma}_I^n \overline{D}_{I-1}}{2h^2}\right)\psi_{I-2}^n + \left(-\frac{\overline{\sigma}_I^n \overline{D}_{I-2}}{8h^2}\right)\psi_{I-1}^n +$$

$$\left(-\frac{\mu_I}{k}+\frac{\overline{\alpha}_I^n}{2}+\frac{\overline{\sigma}_I^n\overline{D}_{I-1}}{2h^2}\right)\psi_I^n+\left(\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-2}}{8h^2}\right)\psi_{I-3}^{n+1}+\left(-\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-1}}{2h^2}\right)\psi_{I-2}^{n+1}+$$
$$\left(-\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-2}}{8h^2}\right)\psi_{I-1}^{n+1}+\left(\frac{\mu_I}{k}+\frac{\overline{\alpha}_I^{n+1}}{2}+\frac{\overline{\sigma}_I^{n+1}\overline{D}_{I-1}}{2h^2}\right)\psi_I^{n+1}=\frac{\overline{W}_I^n+\overline{W}_I^{n+1}}{2}. \quad (8.19)$$

- Para $(i,n)\in\{i^\star,\ldots,I\}\times\{1\}$,

$$\psi_i^1=\overline{f}_i. \quad (8.20)$$

- Para $(i,n)\in\{1,\ldots,i^\star\}\times\{N\}$,

$$\psi_i^N=\overline{g}_i. \quad (8.21)$$

La eficiencia del método directo, en lo que a tiempo de cálculo se refiere, depende fuertemente de la forma en que se lleve a cabo la programación. Por ello, se dedica el capítulo 4 a detallar los aspectos fundamentales del código elaborado en lenguaje MATLAB®. Resulta clave aprovechar las capacidades de vectorización de MATLAB®, ya que hacerlo reduce enormemente el tiempo de computación. Se explica la utilidad del comando `sparse` y se hace uso en repetidas ocasiones, para vectorizar el código, de los comandos `repmat` y `kron`. Las definiciones tanto de la matriz como del segundo miembro del sistema lineal se transcriben desde el código MATLAB® al capítulo 4 de forma literal. Puesto que la programación del esquema par y del esquema impar es muy parecida, nos centramos en el segundo, que es ligeramente más complicado, y además es el que se ha mostrado más efectivo en nuestros ejemplos de prueba.

El capítulo 5 está enteramente dedicado a la descripción del algoritmo iterativo, de nuevo para el esquema impar. La idea es la que se ha venido utilizando en trabajos anteriores a este para problemas similares. Se hace énfasis en que el sistema lineal que se resuelve es el mismo que el descrito para el esquema impar, pero en lugar de resolver un único sistema lineal de orden $(I\times N)\times(I\times N)$ se resuelven varios (que pueden ser muchos) sistemas lineales de orden $\left(\frac{I+1}{2}\right)\times\left(\frac{I+1}{2}\right)$. Si $Q_-=[-1,0)\times[Z_{\text{ini}},Z_{\text{fin}}]$, $Q_+=(0,1]\times[Z_{\text{ini}},Z_{\text{fin}}]$ y $Q_0=\{0\}\times[Z_{\text{ini}},Z_{\text{fin}}]$, y entendiendo "en el interior relativo de Q_0" (es decir, en $\{0\}\times(Z_{\text{ini}},Z_{\text{fin}})$) cuando se dice "en Q_0", el algoritmo iterativo queda descrito por los pasos siguientes:

PASO 0. Se proporciona una semilla en Q_0, que es un conjunto de $N-2$ valores que sustituyan a los desconocidos valores de la solución en los puntos de la malla que caen en Q_0. La elección de una buena semilla tiene incidencia positiva en el número de iteraciones necesarias para la convergencia. Una opción interesante consiste en emplear el método directo con una malla grosera para calcular la semilla mediante interpolación.

PASO 1. Se resuelve hacia delante en Q_+, lo que supone resolver $N-1$ sistemas lineales de orden $\left(\frac{I+1}{2}\right)\times\left(\frac{I+1}{2}\right)$.

PASO 2. Se resuelve hacia atrás en Q_-, lo que supone resolver otros $N-1$ sistemas lineales de orden $\left(\frac{I+1}{2}\right)\times\left(\frac{I+1}{2}\right)$.

PASO 3. Se actualizan los valores en Q_0, haciendo uso de la EFP *(8.13)* restringida a $\{0\}\times$

$(Z_{\text{ini}}, Z_{\text{fin}})$:

$$\alpha(z,0)\psi(z,0) = \sigma(z,0)\frac{\partial^2\psi}{\partial\mu^2}(z,0) + W(z,0), \quad z \in (Z_{\text{ini}}, Z_{\text{fin}}). \tag{8.22}$$

Nótese que una discretización de *(8.22)* con la fórmula centrada estándar para la derivada segunda proporciona una ecuación que relaciona los valores en Q_0 con valores en Q_- y en Q_+, lo que permite la actualización. Con objeto de acelerar la convergencia del algoritmo, estos valores se combinan linealmente con los valores previos en Q_0 haciendo uso de un parámetro de relajación $\omega \in \mathbb{R} \setminus \{0\}$, dando pesos ω y $1-\omega$ a cada una de las partes.

PASO 4. Se comprueba si se ha alcanzado la convergencia con un test que tiende a ser relativo cuando los valores sobre Q_0 son muy grandes y absoluto cuando los valores sobre Q_0 son muy pequeños. El test tiene la ventaja de que puede usarse en cualquier situación, frente a un test puramente relativo, el cual no puede ser usado cuando la solución numérica es 0 o está muy próxima a 0. Si el test se supera, el proceso termina y, en caso contrario, se vuelve al PASO 1.

El capítulo 6 se dedica a efectuar comparaciones entre el método directo y el método iterativo, usando siempre el esquema impar. Específicamente, se compara el tiempo de cálculo requerido por esos dos algoritmos, que, no lo olvidemos, resuelven el mismo problema. Realizar estas comparaciones es natural porque hasta el presente trabajo los métodos que podían encontrarse en la literatura para problemas de este tipo eran iterativos. Se hace notar que los métodos iterativos resuelven en cada iteración $2(N-1)$ sistemas de orden $\left(\frac{I+1}{2}\right) \times \left(\frac{I+1}{2}\right)$, lo cual es mucho menos costoso que resolver un único sistema de orden $(I \times N) \times (I \times N)$, que es lo que se hace con el método directo. La clave está, pues, en el número de iteraciones requeridas para alcanzar la convergencia del método iterativo. Los experimentos numéricos demuestran que casi siempre es ventajoso, en algunas ocasiones muy ventajoso, el empleo del método directo. Las ideas fundamentales son las siguientes:

- Supongamos que el algoritmo iterativo converge en un número de iteraciones que se mantiene acotado por una cantidad que no depende de la malla o que crece moderadamente cuando la malla se refina. Entonces, cuando I y N tienden a infinito, es decir, cuando la malla se refina hasta su límite, el método iterativo acabará antes o después siendo superior al directo, porque el coste de resolver el sistema de orden $(I \times N) \times (I \times N)$ crece con mucha más rapidez que el coste de resolver $2(N-1)$ sistemas de orden $\left(\frac{I+1}{2}\right) \times \left(\frac{I+1}{2}\right)$.

- Si, para una malla dada, el método directo es computacionalmente más barato que el iterativo o no, es una cuestión que depende del número de iteraciones que hacen falta para que el segundo converja. Puesto que no hay estudios teóricos que estimen ese número de iteraciones, se necesitan experimentos numéricos que comparen los tiempos de cálculo.

- El método iterativo puede acelerarse mediante la elección óptima del parámetro de relajación ω. No obstante, no es posible saber a día de hoy, teóricamente, cuál es ese valor óptimo, ni existe tampoco una estimación de tal valor. Los experimentos numéricos demuestran que para un mismo problema el valor óptimo de ω varía con la malla.

También demuestran que el método iterativo diverge (la solución numérica explota) para algunos valores de ω, lo que convierte la elección del parámetro de relajación en un asunto delicado.

- Para que el algoritmo iterativo no se vea perjudicado en la comparación, se hace una búsqueda previa, experimental, del parámetro de relajación óptimo o, mejor dicho, cuasi-óptimo. Además, se usa el método directo con malla grosera para proveer al método de una buena semilla.

- Los experimentos numéricos realizados muestran que, en todas las pruebas realizadas, el método directo es superior al iterativo para el rango de mallas que se usan típicamente. La diferencia entre los tiempos de cálculo es, en algunos casos, muy significativa.

No se cierra la posibilidad a que, en algún caso en el que el producto $I \times N$ sea muy grande, el método iterativo sea el más ventajoso, pero a la vez se recuerda que incluso en una tal situación el método directo tiene la utilidad de poder ser usado para calcular una buena semilla.

Finalmente, en el capítulo 7 se propone un método numérico para resolver el problema *(8.9)–(8.12)*, es decir, el problema dependiente del ángulo acimutal θ. Dicho método permite que las funciones dato W, f y g dependan, además de de z y de μ, de θ, pero su propia naturaleza obliga a que los coeficientes de absorción α y de *scattering* σ, pudiendo naturalmente depender de z y de μ, tengan que ser independientes de θ. En lugar de discretizar un espacio tridimensional de variables (μ, z, θ), se tiene en cuenta la periodicidad de ψ con respecto a θ expresada por la ecuación *(8.12)* para escribirla como su serie de Fourier con respecto a esa variable, es decir,

$$\psi(\mu, z, \theta) = \psi_0(\mu, z) + 2\sum_{k=1}^{\infty} \Big\{ \mathrm{Re}[\psi_k(\mu, z)] \cos(k\theta) \; - \mathrm{Im}[\psi_k(\mu, z)] \sin(k\theta) \Big\},$$

donde los coeficientes ψ_k, $k \in \mathbb{N} \cup \{0\}$, vienen dados por la conocida fórmula

$$\psi_k(\mu, z) = \frac{1}{2\pi} \int_0^{2\pi} \psi(\mu, z, \theta) \; \mathrm{e}^{-ik\theta} \; \mathrm{d}\theta, \tag{8.23}$$

en la que i es la unidad imaginaria. Se deduce de *(8.23)* que es posible obtener una buena aproximación de ψ por truncamiento de la serie cuando dicha serie converge rápidamente, en caso de que tengamos una forma de calcular los coeficientes ψ_k. Se prueba que ψ_k es la solución de la EDP

$$\mu \frac{\partial \psi_k}{\partial z} + \left(\alpha + \frac{k^2 \sigma}{1 - \mu^2} \right) \; \psi_k = \sigma \frac{\partial}{\partial \mu} \left[(1 - \mu^2) \; \frac{\partial \psi_k}{\partial \mu} \right] + W_k \tag{8.24}$$

con las condiciones de cierre

$$\psi_k(\mu, Z_{\mathrm{ini}}) = f_k(\mu) \text{ para } \mu \in (0, 1], \tag{8.25}$$
$$\psi_k(\mu, Z_{\mathrm{fin}}) = g_k(\mu) \text{ para } \mu \in [-1, 0), \tag{8.26}$$

donde hay que entender que el subíndice k indica el k-ésimo coeficiente de Fourier con respecto a θ también para W, f y g. El problema *(8.24)–(8.26)* no entra directamente dentro del marco del problema nuclear *(8.13)–(8.15)* porque cuando $k \geq 1$ el coeficiente de absorción $\alpha + \frac{k^2\sigma}{1-\mu^2}$

es singular en $\mu = \pm 1$. Es sin embargo sencillo emplear el mismo esquema impar descrito en el capítulo 3 para el problema nuclear, una vez que dicho esquema se modifica ligeramente para poder tratar esas singularidades. El método descrito, programado empleando MATLAB®, ofrece resultados numéricos que muestran de nuevo orden 2 de convergencia.

La memoria se termina con la exposición de las conclusiones y de algunas ideas para trabajos futuros.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Bibliography

[1] Aziz, A. Kadir; French, Donald A.; Jensen, Soren and Kellogg, Royal Bruce. Origins, analysis and numerical approximation of a forward-backward parabolic problem, *Mathematical Modelling and Numerical Analysis* **33** no. 5 (1999) 895-922.

[2] Bartine, David Elliott. *Low-energy Electron Transport by the Method of Discrete Ordinates.* Ph.D thesis, University of Missouri-Rolla (now Missouri University of Science and Technology), 1971.

[3] Beals, Richard. On an equation of mixed type from electron scattering theory, *Journal of Mathematical Analysis and Applications* **58** no. 1 (1977) 32-45.

[4] Beals, Richard. Indefinite Sturm-Liouville problems and half-range completeness, *Journal of Differential Equation* **56** no. 3 (1985) 391-407.

[5] Davis, Timothy A. (2006) *Direct Methods for Sparse Linear Systems.* SIAM, Philadelphia, PA.

[6] Degond, Pierre and Mas-Gallic, Sylvie. Existence of solutions and diffusion approximation for a model Fokker-Planck equation, *Transport Theory and Statistical Physics* **16** no. 4-6 (1987) 589-636.

[7] Dieudonné, Jean (1969) *Foundations of Modern Analysis.* Academic Press, New York and London. (Third (enlarged and corrected) printing).

[8] Dziuk, Gerhard. Finite elements for the Beltrami operator on arbitrary surfaces, *Partial Differential Equations and Calculus Of Variations* **1357** (Lecture Notes in Math.,) (1988) 142-155.

[9] Dziuk, Gerhard and Demlow, Alan. An adaptive finite element method for the Laplace-Beltrami operator on implicitly defined surfaces, *SIAM Journal on Numerical Analysis* **45** no. 1 (2007) 421-442.

[10] Dziuk, Gerhard and Elliott, Charles M. Finite element methods for surface PDEs, *Acta Numerica* **22** (2013) 289-396.

[11] Epstein, Charles L. and Mazzeo, Rafe (2013) *Degenerate Diffusion Operators Arising in Population Biology.* Princeton University Press, Princeton, NJ.

[12] Fisch, Nathaniel Joseph and Kruskal, Martin David. Separating variables in two-way diffusion equations, *Journal of Mathematical Physics* **21** no. 4 (1980) 740-750.

[13] Frank, Martin; Hensel, Hartmut and Klar, Axel. A fast and accurate moment method for the Fokker-Planck equation and applications to electron radiotheraphy, *SIAM Journal on Applied Mathematics* **67** no. 2 (2006/2007) 582-603.

[14] Franklin, Joel N. and Rodemich, Eugene R. Numerical analysis of an elliptic-parabolic partial differential, *SIAM Journal on Numerical Analysis* **5** no. 4 (1968) 680-716.

[15] Gorbikov, Sergey Pavlovich and Melnikov, Viktor Fedorovich. The numerical solution of the Fokker-Planck equation for modeling of particle distribution in solar magnetic traps, *Mathematical Modeling* **19** no. 2 (2007) 112-122.

[16] Grisvard, Pierre (1985) *Elliptic Problems in Nonsmooth Domains*. University of Nice, Pitman Publishing, Marshfield, Massachusetts.

[17] Hämmerlin, Günther; Hoffmann, Karl-Heinz (1991) *Numerical Mathematics*. Springer, New York, NY.

[18] Han, Houde and Yin, Dondgsheng. A non-overlap domain decomposition method for the forward-backward heat equation, *Journal of Computational and Applied Mathematics* **159** no. 1 (2003) 35-44.

[19] Hensel, Hartmut; Iza-Teran, Rodrigo and Siedow, Norbert. Deterministic model for dose calculation in photon radiotheraphy, *Physics in Medicine and Biology* **51** no. 3 (2006) 675-693.

[20] Herty, Michael; Jörres, Christian and Sandjo, Albert Nana. Optimization of a model Fokker-Planck equation, *Kinetic an Related Models* **5** no. 3 (2012) 485-503.

[21] Kim, Arnold D. and Tranquilli, Paul. Numerical solution of the Fokker-Planck equation with variable coefficients, *Journal of Quantitative Spectroscopy and Radiative Transfer* **109** no. 5 (2008) 727-740.

[22] Klaus, Martin; van der Mee, Cornelis Victor Maria and Protopopescu, Vladimir. Half-range solutions of indefinite Sturm-Liouville problems, *Journal of Functional Analysis* **70** no. 2 (1987) 254-288.

[23] Klose, Alexander D. and Larsen, Edward W. Light transport in biological tissue based on the simplified spherical harmonics equations, *Journal of Computational Physics* **220** no. 1 (2006) 441-470.

[24] Larsen, Edward W. and Morel, Jim A. (2010) Advances in Discrete-Ordinates methodology. In *Nuclear Computational Science: A Century in Review*, Yousry Azmy and Enrico Sartori (editors), pages 1-84. Springer, Dordretch [etc.].

[25] López Pouso, Óscar and Jumaniyazov, Nizomjon. Numerical experiments with the

Fokker-Planck equation in 1D slab geometry, *Journal of Computational and Theoretical Transport* **45**, no. 3 (2016) 184-201.

[26] López Pouso, Óscar and Jumaniyazov, Nizomjon. Direct versus iterative methods for forward-backward diffusion equations. Numerical comparisons on a particular transport kinetic model, *submitted*.

[27] Morel, Jim E. An improved Fokker-Planck angular differencing scheme, *Nuclear Science and Engineering* **89** no. 2 (1985) 131-136.

[28] Sheng, Qiwei and Han, Weimin. Well posedness of the Fokker-Planck equation in a scattering process, *Journal of Mathematical Analysis and Applications* **406** no. 2 (2013) 531-536.

[29] Stein, Daniel and Bernstein, Ira B. Boundary value problem involving a simple Fokker-Planck equation, *Physics of Fluids* **19** no. 6 (1976) 811-814.

[30] Sun, Jie. Numerical schemes for the forward-backward heat equation, *International Journal of Computer Mathematics* **87** no. 3 (2010) 552-564.

[31] Tolstov, Georgi Povlovich (1976) *Fourier Series.* Dover Publications, New York. (Revised reprint of the Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962 edition.)

[32] Vanaja, Venkataraman. Numerical solution of a simple Fokker-Planck equation, *Applied Numerical Mathematics* **9** no. 6 (1992) 533-540.

[33] Vanaja, Venkataraman; Kellogg, Royal Bruce. Iterative methods for a forward-backward heat equation, *SIAM Journal on Numerical Analysis* **27** no. 3 (1990) 622-635.

[34] Ziegler, James F. Stopping of energetic light ions in elemental matter, *Journal of Applied Physics* **85** no. 3 (1999) 1249-1272.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA


USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA